Technische Universität Dresden

An Integrated Approach to Modeling German Prosody

Dr.-Ing. Hansjörg Mixdorff

Der Fakultät Elektrotechnik und Informationstechnik
der Technischen Universität Dresden

zur Erlangung des akademischen Grades

Doktor-Ingenieur habilitatus
(Dr.-Ing. habil.)
vorgelegte Habilitationsschrift
Zusammensetzung des Habilitationsausschusses:

Vorsitzender:                                      Prof. Dr.-Ing. Karl-Heinz Gonschorek
1. Gutachter:                                      Prof. Dr.-Ing. habil. Rüdiger Hoffmann
2. Gutachter:                                      Prof. Dr.-Ing. Hiroya Fujisaki
3. Gutachter:                                      Prof. Dr. John Local

Tag der Einreichung:                               9.11.2001
Tag der wissenschaftlichen Aussprache:             25.7.2002
For Noi
Abstract

This thesis documents the past five years of work in prosody research performed by the author. The thesis starts from the quantitative model of German intonation, which the author developed in his D.Eng. thesis, and which uses the Fujisaki model for parametrizing $F_0$ contours. The symbolic representation of German intonation adopted is based on the concept of "tone switches" first introduced by Isacenko and further developed into a theory of basic intonational units, called 'intonemes' by Stock et al. Intonemes are categorized by the type of tone switch, that is, distinctive transitions in the $F_0$ contours which occur at accented syllables, with which they are associated.

When applied to Text-to-Speech synthesis, the quantitative model (henceforth called MFGI, Mixdorf-Fujisaki model of German Intonation) produces the $F_0$ contour for a given sentence following a two-stage process: (1) Symbolic Prosody Generation, (2) $F_0$ Contour Generation.

First, by applying accentuation and phrasing rules originally developed by Stock, but further refined by the author, phrase boundaries and accented syllables are determined. These lead to the appropriate sequence of intonemes underlying the utterance to be generated, as well as the locations and depths of phrase boundaries, i.e. the symbolic prosody. This information is then used in a second step, the $F_0$ contour generation, again by applying a set of rules, to derive the amplitudes of phrase and accent commands and the fine timing of these commands with respect to the phones in the utterance. The $F_0$ contour proper is then straightforwardly computed by the Fujisaki model.

MFGI was implemented during the course of a two-year DFG research project, integrated into the Dresden TTS system (DRESS) and evaluated in a series of perceptual studies.

These studies concerned the comparison of several $F_0$ contour generation modules tested within the DRESS framework, as well as the comparison between DRESS and other TTS systems for German.

Experimental results showed that, although the model yielded a higher naturalness compared with other approaches, the imperfections of the TTS system's rule-based duration module impaired the speech quality produced. Furthermore, the obvious deficiencies compared with natural speech were still paramount.

Based on these observations, the author decided to investigate whether an integrated approach to modeling prosody producing durations and $F_0$ in parallel would be appropriate to yield prosodically more coherent synthetic speech. The syllable was chosen as the modeling unit for anchoring the prosodic features.

In a series of preliminary studies the relationship between the durational and $F_0$ contours was examined. These showed that, inter alia, (1) the timing of intonational events is influenced significantly by the phonetic structure of the accented syllables these events are connected to, (2) the perceived prominence of syllables is closely related to the accent command amplitude $AA$ and the duration of the respective syllable, (3) the focal structure of an utterance is closely tied to the $F_0$ contour, whereas boundaries are consistently signaled by lengthening of pre-boundary syllables.

In contrast to the original rule-based approach, a neural network trained with data from a speech corpus was implemented. Comparing the performance of the neural network with linear regression models as a baseline for predicting individual parameters, however, did not show a consistent advantage of the joint prediction implemented in the integrated model.

An extensive evaluation was performed for assessing the perceptual quality of the integrated model. An important paradigm adopted in the evaluation was to employ resynthesized stimuli which were created by prosodic degrading of natural speech. By applying this technique a reference matrix of segmentally high-quality stimuli was yielded which were defined by their
distance from the original speech in terms of the de-correlation in the $F_0$ and duration domains. The main outcomes of the perceptual study were as follows: (1) Subjects were far more sensitive to deviations in the duration contours than to deviations in the $F_0$ contours, (2) the integrated model was perceived more natural than degraded stimuli of comparable correlations between observed and predicted parameters, (3) the integrated model performed less acceptably on sentences from outside the corpus from which it was trained, (4) the integrated approach outperformed the original rule-based model mainly in terms of the accuracy of its prediction of duration, not in terms of the quality of the $F_0$ contour. A major conclusion from the latter result is that the input information extracted from text is insufficient for predicting relative constituent prominence, a major correlate of meaning.
Contents

0.1 Table of Abbreviations ................................................... xi
0.2 List of Symbols ............................................................ xi
0.3 Typography ................................................................. xii
0.4 SAMPA Reference Table .................................................. xii

1 Introduction ................................................................. 1

2 Prosodic Features of Speech ........................................... 5
  2.1 The Term ‘Prosody’ Employed in this Thesis .................... 6
  2.2 Speech Production ....................................................... 8
    2.2.1 Fundamental Frequency (F0) ................................. 8
    2.2.2 Syllable Duration ............................................ 10
  2.3 Speech Perception ...................................................... 13
    2.3.1 Perception of Intonation .................................. 13
    2.3.2 Perception of Rhythm ....................................... 14
  2.4 Measuring Prosodic Features ....................................... 15
  2.5 Information Encoded in Prosodic Features ...................... 16
    2.5.1 Functions of Prosody ....................................... 16
    2.5.2 Word accent ............................................... 17
    2.5.3 Accent, Stress, Pitch Accent and Stress Accent .......... 18
    2.5.4 Accent Group, Prosodic Phrase, Sentence Accent ........ 19
    2.5.5 Focus ...................................................... 22
    2.5.6 Sentence Mode ............................................. 23
    2.5.7 Declination ............................................... 23
    2.5.8 Location of F0 peaks ...................................... 23
    2.5.9 Conclusions ............................................... 25
  2.6 Models of Prosody .................................................... 26
    2.6.1 Introduction ............................................... 26
    2.6.2 Duration Models .......................................... 27
    2.6.3 Intonation Models ......................................... 37
  2.7 The quantitative intonation model MFGI ......................... 51
    2.7.1 Linguistic Background of MFGI ............................ 51
    2.7.2 The Fujisaki Model ........................................ 54
    2.7.3 Components of MFGI ....................................... 56

3 Perceptual Evaluation of MFGI ....................................... 62
  3.1 Introduction .......................................................... 63
  3.2 Evaluation Designs Applied ...................................... 63
    3.2.1 Pair Comparison of Synthetic Stimuli .................... 64
    3.2.2 Ranking TTS systems ...................................... 66
3.3 \(F_0\) Control within DRESS ................................................................. 66
  3.3.1 The Experimental TTS system .................................................. 68
  3.3.2 The Corpus Used ........................................................................ 69
  3.3.3 Subjects ...................................................................................... 69
  3.3.4 Intelligibility .............................................................................. 69
  3.3.5 Perceived Naturalness (Grading) .................................................. 69
  3.3.6 Perception of Intended Accents .................................................. 70
  3.3.7 Perceived Naturalness in a Pair-Comparison .................................. 72
  3.3.8 Fully Synthetic against Reference Examples .................................. 73
3.4 Comparison of Complete TTS Systems .............................................. 78
  3.4.1 Pair-Comparison of Isolated Sentences ........................................ 79
  3.4.2 System Comparison ................................................................... 88
  3.4.3 Discussion .................................................................................. 92
3.5 Discussion and Conclusions ................................................................. 93

4 \(F_0\) and Duration - Preliminary Studies ............................................... 94
  4.1 Introduction .................................................................................... 95
  4.2 The Influence of the Syllable Structure on the \(F_0\) contour ................. 95
    4.2.1 Introduction ............................................................................ 95
    4.2.2 Speech Material ..................................................................... 95
    4.2.3 Method of Analysis ................................................................ 96
    4.2.4 Results of Analysis ................................................................ 96
    4.2.5 Perceptual Evaluation ............................................................. 102
    4.2.6 Discussion and Conclusions .................................................... 102
  4.3 Perception of Prominence ................................................................. 102
    4.3.1 Introduction ............................................................................ 102
    4.3.2 Speech Material and Method of Analysis .................................. 103
    4.3.3 Results of Analysis ................................................................ 104
    4.3.4 Discussion and Conclusions .................................................... 106
  4.4 Focus, Sentence Mode and Phrase Boundary Location ....................... 107
    4.4.1 Introduction ............................................................................ 107
    4.4.2 Speech Material and Method of Analysis .................................. 108
    4.4.3 Results of Analysis ................................................................ 110
    4.4.4 Discussion and Conclusions .................................................... 112
  4.5 Summary of Preliminary Studies ....................................................... 112

5 An Integrated Approach to Modeling Prosody ........................................ 114
  5.1 Introduction .................................................................................... 115
  5.2 Selecting a Database ..................................................................... 115
  5.3 Parameter Extraction and Organization ........................................... 117
    5.3.1 Organizing the Database ......................................................... 117
    5.3.2 Fujisaki Model Parameter Estimation ....................................... 117
  5.4 Statistical Analysis of the Prosodic Database .................................... 124
    5.4.1 Introduction ............................................................................ 124
    5.4.2 Analysis of Syllable Durations ............................................... 125
    5.4.3 Analysis of Fujisaki Control Parameters ................................... 136
    5.4.4 Pauses ................................................................................... 146
    5.4.5 Fujisaki Model Parameter vs. ToBI Labels .............................. 146
  5.5 Properties of the IGM .................................................................. 151
5.6 Training and Testing the IGM
5.7 Discussion and Conclusions

6 Perceptual Evaluation of the Integrated Model
6.1 Introduction
6.2 Experiment Design
6.3 Selection of Test Sentences for the Evaluation
6.4 Creating a Matrix of Reference Stimuli
6.5 Experiment 1: The IGM in a Matrix of Reference Stimuli
  6.5.1 Introduction
  6.5.2 Setting
  6.5.3 Results
  6.5.4 Discussion
6.6 Experiment 2: Sentences not Pertaining to the News Corpus
  6.6.1 Introduction
  6.6.2 Setting
  6.6.3 Results
  6.6.4 Discussion
6.7 Experiment 3: The IGM Compared with the Original RBM
  6.7.1 Introduction
  6.7.2 Setting
  6.7.3 Results
  6.7.4 Discussion
6.8 Experiment 4: IGM vs. RBM (Diphone Synthesis), Intensity
  6.8.1 Introduction
  6.8.2 Setting
  6.8.3 Results
  6.8.4 Discussion
6.9 Discussion and Conclusions

7 Discussion and Conclusions

A Glossary of Special Terms Used in the Thesis
B Corpus of Sentences used in Perception Experiments
C Acknowledgements
D Curriculum Vitae
### 0.1 Table of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BI</td>
<td>Break Index</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and Regression Tree</td>
</tr>
<tr>
<td>DRESS</td>
<td>TU Dresden TTS system</td>
</tr>
<tr>
<td>FFNN</td>
<td>Feed-forward Neural Network</td>
</tr>
<tr>
<td>GLM</td>
<td>General Linear Model</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>HFC</td>
<td>High Frequency Contour (see Glossary, Appendix A)</td>
</tr>
<tr>
<td>IGM</td>
<td>Integrated Model of Prosody (see Glossary, Appendix A)</td>
</tr>
<tr>
<td>IPA</td>
<td>International Phonetic Alphabet</td>
</tr>
<tr>
<td>IPCG</td>
<td>Inter-Perceptual Center Group</td>
</tr>
<tr>
<td>LFC</td>
<td>Low Frequency Contour (see Glossary, Appendix A)</td>
</tr>
<tr>
<td>MARS</td>
<td>Multi-adaptive Regressive Splines</td>
</tr>
<tr>
<td>MFGI</td>
<td>Mix dorff-Fujisaki model of German Intonation (see Glossary, Appendix A)</td>
</tr>
<tr>
<td>MLP</td>
<td>Multi-layer Perceptron</td>
</tr>
<tr>
<td>PSOLA</td>
<td>Pitch Synchronous Overlap-Add</td>
</tr>
<tr>
<td>PURRE</td>
<td>Prosody Unveiling Restricted Representation</td>
</tr>
<tr>
<td>RBM</td>
<td>Rule-based Model of Prosody (see Glossary, Appendix A)</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root mean square error</td>
</tr>
<tr>
<td>SAMPA</td>
<td>Speech Assessment Methods Phonetic Alphabet</td>
</tr>
<tr>
<td>TOBI</td>
<td>Tones and Break Indices</td>
</tr>
<tr>
<td>TTS</td>
<td>Text-to-Speech</td>
</tr>
<tr>
<td>VCG</td>
<td>Verb-Complement Group</td>
</tr>
</tbody>
</table>

### 0.2 List of Symbols

- $F_0$: Fundamental frequency
- $T_0$: Fundamental period
- $z_i$: $z$-score (with respect to segment durations)
- $\text{dur}_i$: duration of a segmental token $i$
- $\mu_i$: mean duration of all tokens pertaining to the same segment class as token $i$
- $\sigma_i$: standard deviation of duration of all tokens pertaining to the same segment class as token $i$
- $A_p$: phrase command magnitude
- $T_0$: phrase command onset time
- $A_a$: accent command amplitude
- $T_1$: accent command onset time
- $T_2$: accent command offset time
- $\uparrow_E$: Rising tone-switch early in the accent syllable
- $\uparrow_L$: Rising tone-switch late in the accent syllable
- $\downarrow_E$: Falling tone-switch early in the accent syllable
- $\downarrow_L$: Falling tone-switch late in the accent syllable
- $I\downarrow$: Information intone
- $C\uparrow$: Contact intone
- $N\uparrow$: Non-terminal intone
- $B_{\text{cat}} \uparrow$: Boundary Tone, concatenated type
- $B_{\text{noncat}} \uparrow$: Boundary Tone, non-concatenated type
- $\rho_{\text{dur}}$: Correlation between measured and predicted syllable durations
- $\rho_{F_0}$: Correlation between measured and predicted $F_0$ contours
0.3 Typography

*Italic* s are generally used to indicate English translations of German exemplary texts. *Syllables* which are locations of word accents are set in **bold** type. In Figures, words which are the location of a narrow focus are CAPITALIZED. German ‘Umlaut’ characters in plots of prosodic parameters are printed using double quotes, i.e. ”a” for ‘ä’, for instance.

0.4 SAMPA Reference Table

Throughout this thesis the Speech Assessment Methods Phonetic Alphabet (SAMPA) transcription will be employed. The following is a list of IPA symbols for German with their SAMPA correspondences and two examples each. The graphemic part of the words pertaining to the phoneme is capitalized.

<table>
<thead>
<tr>
<th>IPA</th>
<th>SAMPA</th>
<th>Examples(orthographic)</th>
<th>IPA</th>
<th>SAMPA</th>
<th>Examples(orthographic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>_</td>
<td>?</td>
<td>before syllable-initial vowel</td>
<td>a</td>
<td>I</td>
<td>Liebe, halLe</td>
</tr>
<tr>
<td>?</td>
<td>R</td>
<td>Riese, kRaut</td>
<td>p</td>
<td>6</td>
<td>opER, deR</td>
</tr>
<tr>
<td>b</td>
<td>J</td>
<td>Jetzt, Jagd</td>
<td>t</td>
<td>aI</td>
<td>Elns, kAIser</td>
</tr>
<tr>
<td>d</td>
<td>?Y</td>
<td>Auflerung, nEU</td>
<td>k</td>
<td>aO</td>
<td>AUf, schAU</td>
</tr>
<tr>
<td>g</td>
<td>ø</td>
<td>schEn, bEsagt</td>
<td>f</td>
<td>i</td>
<td>Igel, bIten</td>
</tr>
<tr>
<td>v</td>
<td>I</td>
<td>In, bitten</td>
<td>s</td>
<td>y</td>
<td>Ubung, hUten</td>
</tr>
<tr>
<td>z</td>
<td>Y</td>
<td>Ypsilon, hUten</td>
<td>j</td>
<td>e</td>
<td>bEten, schnEE</td>
</tr>
<tr>
<td>z</td>
<td>e</td>
<td>bEten, schnEE</td>
<td>5</td>
<td>ë</td>
<td>bEtten, gAst</td>
</tr>
<tr>
<td>x</td>
<td>e</td>
<td>Asen, gebIase</td>
<td>x</td>
<td>ë</td>
<td>Asen, gebIase</td>
</tr>
<tr>
<td>C</td>
<td>ò</td>
<td>òf</td>
<td>h</td>
<td>9</td>
<td>Ofen, kOmen</td>
</tr>
<tr>
<td>pf</td>
<td>u</td>
<td>bUHen, gUt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ts</td>
<td>u</td>
<td>ëf</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tf</td>
<td>ò</td>
<td>òf</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>ò</td>
<td>òf</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>a</td>
<td>wAr, wAHR</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| y   | a     | An, kAm
Chapter 1

Introduction

In this chapter the motivation and aims of the present work are briefly discussed. An introduction is given as to the scope and structure of the thesis.
CHAPTER 1. INTRODUCTION

This work starts from my D.Eng. thesis “Intonation patterns of German” [Mix98] which developed the outline of a quantitative model of German intonation. During the course of a two-years’ project (1996 – 1998) funded by the Deutsche Forschungsgemeinschaft this model was incorporated into the high-quality TTS system developed at TU Dresden and subject to a larger scale perceptual evaluation. One of the main outcomes of this project was that though the intonation model produced quite naturally sounding $F_0$ contours - which placed the Dresden system in the top group of German TTS systems available at that time - its performance was flawed by the imperfect modeling of segment durations used. This led me to the conclusion that as long as we treat the prosodic features $F_0$ and segment duration\(^1\) separately, we might end up with TTS systems of suboptimal prosodic quality.

From the way speech is produced we could be easily tempted to believe that we can tackle the problem of modeling $F_0$ and duration separately, since the former is mainly controlled by laryngeal muscles and the latter by the movements of the articulators in the vocal tract, i.e. the tongue, the velum, the jaw, and the lips. However, since speech is perceived holistically (ganzeheitlich), the separation of intonation and speech rhythm\(^2\) is at least questionable. Perception experiments using synthetic speech have shown that, for instance, syllable prominence can be increased by stretching the sounds of the syllable, as well as by increasing the amplitude of $F_0$ movement assigned to it. We should also keep in mind that prosodic features are speaker-individual, and that building intonation models and duration models, either from entirely different sets of speech data, or by using rules derived from different speakers contradicts this fact.

We make use of prosody partly consciously, partly unconsciously for structuring and modifying the meaning of an utterance by highlighting or dimming its constituents. In the study of prosody, however, we always need to keep in mind that we are looking only at one means of modifying the contents of a verbal message, and that there are others such as changing the syntax (the word order, for instance) or using certain morphological items such as interrogatives which might be applied more frequently. Hence, instances where prosody truly is the one single means of contrasting two utterances (also called ‘minimal pairs’) are seldom found in conversational speech and even less so in read speech.

Still, intonation and speech rhythm accompany every speech act, even nonsense utterances or babble\itude speech. We readily identify speech as sounding unnatural, if suprasegmental features do not behave according to the mental patterns we acquired during the forming of our own speech capabilities. This makes synthetic speech readily identifiable, even if the segments from which it is build are taken from a natural source and do not exhibit strong distortion on the segmental level. Furthermore, the lack of naturalness in prosody greatly reduces the intelligibility of synthetic speech as has been shown in identification tasks via a telephone line.

The importance of these observations has triggered broad research activities towards understanding the way how prosody works and how to adequately predict the prosodic features of

\(^1\)I am aware of the fact that $F_0$ and duration - strictly speaking the variation of segment duration for a particular phoneme class against a defined mean value - are not the only prosodic features. Intensity and voice quality, for instance, are also prosodic features co-occurring in the speech signal, but (1) they are of less importance for the quality of synthetic speech, and (2) are no direct correlates of linguistic contents. Still for brevity throughout this work I will use the term ‘prosody’ for denoting intonation, and the variation of segment duration as explained.

\(^2\)Throughout this thesis for better distinction the terms rhythm and intonation will be employed when referring to the principles governing the temporal and melodic organization of speech whereas the terms duration and fundamental frequency ($F_0$) will be used when referring to the measurable correlates of rhythm and intonation in the speech signal. As a consequence, the term ‘$F_0$ model’ refers to an algorithm for generating $F_0$ contours within a TTS system, whereas intonation models are models describing the phonological intonation system of a particular language. Although $F_0$ models are usually based on some kind of intonation model, the latter do not necessarily possess a back-end for calculating $F_0$ contours, but can be well confined to some kind of abstract representation.
speech for speech synthesis. Although the research area is a relatively young one, the growing number of articles in sessions on prosody at international speech conferences such as ICSLP or Eurospeech shows a fast developing interest in the field. As modern, instrument-powered prosody research itself is an on-going task which in its course often raises new questions than answering old ones, the aim of the current work is two-fold: to present an overview of the current state of prosody research, and to document the particular strain of research activities pursued by the author which does not necessarily reflect the main stream.

A main issue in prosody research is the selection of an appropriate corpus on which a study can be based. Although statistical methods such as neural networks or HMM, which require large labeled speech databases, have proved relatively successful for predicting prosodic features of speech, they do not necessarily widened our understanding as to how prosody works. As long as we confine ourselves, say, to the synthesis of declarative sentences, which may be a legitimate limitation, we can well work on read speech corpora which are readily available.

If we, however, wish to look into the contrastive functions of prosody, constructing a special corpus is necessary which reflects the functions we wish to study. The analysis of spontaneous speech might be another choice, though we might end up with a lot of data without pin-pointing the linguistic functions or the contrasts we wish to examine.

Once we have chosen an appropriate corpus we find ourselves confronted with the question where to look for the information we wish to identify in the speech signal. Although extracting the $F_0$ contour of an utterance has become a trivial process, the interpretation of a contour with its sometimes spurious variations is a difficult endeavor. Recently, the ToBI (tones and break indices) system has become widely adopted for labeling $F_0$ contours by assigning high and low tone labels to major 'tonal targets' at accented syllables in an $F_0$ contour, and associating prosodic boundaries indices indicating boundary strength. Apart from inevitable intra-labeler and inter-labeler-inconsistencies, the system does not provide any means of quantifying the height of its tonal targets, except for the assignments of 'downstepped' labels to less prominent accents. Resynthesis of an $F_0$ contour from ToBI-labels only is therefore not possible. Furthermore, the labels themselves do not indicate any linguistic function of the accents to which they are assigned. For this reason, a section of this work is dedicated to the comparison between the quantitative approach adopted by the author and a ToBI-style representation.

This thesis is structured as follows: Chapter 2 starts off with Section 2.1 giving an overview of prosody, such as definitions in the literature and which notion of prosody will be followed in this work. Section 2.2 is dedicated to the mechanism of speech production involved in producing $F_0$ and speech segments and syllables. Section 2.3 discusses findings as to the perception of intonation and speech rhythm. After a short note on measuring prosodic features in Section 2.4, Section 2.5 giving an overview of what types of information is coded in the prosodic features of speech, Section 2.6 discusses existing intonation and duration models, i.e. models which aim at predicting prosodic features from a set of input information. The intonation model developed in my D.Eng. thesis, the Mixdorf-Fujisaki Model of German Intonation, short MFGI, will then be revisited and we shall point out the relationship between the intonational units defined herein and durational units.

Chapter 3 documents a series of perceptual experiments for evaluating MFGI which were performed during the course of the above-mentioned DFG-project on $F_0$ control in TTS. We document the perceptual experiments proper, laying emphasis on the results indicating the need for an integrated model of prosody. In Chapter 4 we document three preliminary studies of interactions between duration and the $F_0$ contour, i.e. (1) the influence of syllable structure on accent command timing, (2) perceived syllable prominence in terms of the accent command amplitude and normalized syllable duration, (3) the influence of linguistic information such as focus, sentence mode and segmentation on the duration and $F_0$ contours.
Chapter 5 describes the design and development of the integrated prosody model. Chapter 6 deals with perception experiments comparing the prosodic quality of the original and the new integrated approach. Chapter 7 completes this work with a discussion and conclusions.
Chapter 2

Prosodic Features of Speech

Abstract
This chapter begins with an overview of prosody, discusses definitions in the relevant literature and which notion of prosody will be followed in this work. The second section of the chapter is dedicated to the mechanism of speech production and explains the physiological background of prosodic speech features of speech and their perception. This is followed by a brief discussion of measurement methods. The last section of this chapter gives an overview of information which is coded in the prosodic features of speech, and discusses existing duration and intonation models, with special focus on approaches which make an attempt to view both features jointly. The intonation model developed in my D.Eng. thesis will then be revisited and discussed with respect to its implementation for TTS.
2.1 The Term ‘Prosody’ Employed in this Thesis

In the study of the relevant literature on intonation and speech rhythm one is confronted with a multitude of definitions and terms. These are seldom standardized and usually denote special manifestations of speech features associated with intonation and rhythm. The term ‘prosody’ obviously belongs to the most vaguely defined terms and is often employed with meanings quite remote from its original notion. As a consequence, we feel the need to discuss the origins of this term and its use in the literature. According to the German Duden Lexicon [DR97], in classical metrics ‘prosody’ denoted ‘the theory of pitch and syllable quantity’, whereas nowadays it is mostly understood as ‘the theory of the metric-rhythmic use of speech’ which is quite an unclear definition. More specific information can be found in a basic lexicon of linguistics [ULr75], where ‘prosody’ is defined as denoting features of speech, which determine the pitch, duration (quantity) and loudness (quality) in single speech sounds, and accent and rhythm in sequences of sounds. This can be interpreted in a way that ‘prosody’ denotes certain features of speech, whose measurable correlates can be found in fundamental frequency, segment duration and intensity. Furthermore, it can be derived that prosodic features occur on the phoneme level as well as on the level of syllables, words and phrases, or any other relevant ‘prosodic units’1. The rather singular notion of ‘prosody’ employed by Artemov [Art78] strictly separates ‘prosody’ from ‘intonation’, denies the former a function modifying meaning and assigns it only formal character in the ‘correct pronunciation of syllables’. Since Artemov claims that the acoustic correlates for both prosody and intonation be the same, his definition of these terms remains vague.

Trubetzkoy defines the prosodic features of speech as a category of acoustic features which are rhythmic-melodic in nature and concern phonemes or phoneme sequences. These features are connected with the syllable, or strictly speaking with a certain part of the syllable, called the ‘syllable carrier’. According to Trubetzkoy there are two classes of prosodic features: (1) features concerning the influence of different phonemes on the F0 contour, also known as ‘microprosody’ (see, for instance [Mix98, p. 62]), (2) features facilitating a rhythmic-melodic differentiation, such as the prominence-lending peaks in the F0 contour, also called ‘accents’. In this context Trubetzkoy considers ‘intonation’ as having a sentence-differentiating function, for distinguishing between terminal and non-terminal utterances, for instance. In more recent literature the term ‘prosody’ is used as a category for a wide range of speech features which transcends the mere articulation of speech sounds. These have in common that they do not provide a clear definition of what they understand by ‘prosody’. Some of these publications concern features such as accent and prominence, others deal with speakers’ emotion. Besides, the relevant domains in which prosodic features are observed vary between authors, and may concern single phonemes, syllables or even entire syntactic phrases. In some studies, the term ‘prosody’ is even employed as a synonym for intonation (see, for instance, [FSH95, Bru95, TI94a, TI94b]).

An ITG (Informationstechnische Gesellschaft, IT society of Germany) [ITG96] recommendation on terminology in speech acoustics describes ‘prosody’ as being “a generic term for intonation and accentuation, as well as for the rhythmic arrangement of speech utterances”, hence stressing that speech melody and rhythm are equally concerned.

Ladd and Cutler [CL83] broadly characterize approaches to prosody as either being ‘concrete’ — i.e., defining prosody from measurements of the speech signal and in terms of acoustic parameters such as pitch, duration and intensity — or ‘abstract’ — deriving prosody from a linguistic point of view in terms of the phonological organization on various levels that eventually leads to observable utterances. This dichotomy applies to prosody models in general as will be

---

1In analogy to the terms ‘phoneme’ or ‘grapheme’, these prosodic units are sometimes called ‘prosodeme’ [Tru77].
shown in Section 2.6.

In order to get a complete picture of prosody, however, one has to look at 'both sides of the coin', especially when the improvement of speech synthesis is concerned. Performing measurements without knowing what to look for, i.e., what information is potentially coded in an utterance, is questionable, simply because of the great variability of the speech signal. As Fujisaki [Fuj97] points out, though measurements of the observable phenomena are important, they are more powerful when they can be generalized in a theory, a method he calls abduction. On the other hand, theories must be validated by making predictions of the observable phenomena (deduction). As a consequence, Fujisaki proposes that "prosody is the systematic organization of various linguistic units into an utterance or a coherent group of utterances in the process of speech production." The approach to prosody developed by the author of this thesis follows this proposed kind of 'two-sided' concept. By applying linguistic knowledge to the analysis of the natural speech signal, the relationships between the information underlying an utterance and the observed prosodic phenomena are established. The concepts developed are then evaluated by applying them to speech synthesis.

![Speech waveform](image)

Figure 2.1: The key to figures showing speech data as used in the current study. From top to bottom the speech waveform, the extracted $F_0$ contour, the contour produced with the Fujisaki-model and the corresponding accent commands are displayed. Word boundaries are marked by the vertical dotted lines.

Before we proceed we need to define the prosodic features we shall examine. These are speech features which affect temporal domains above the segments (syllables, phonemes) and hence have 'suprasegmental' character. The quantifiable acoustic correlates of the most important features are:

**The fundamental frequency contour** $F_0(t)$ which corresponds to the periodicity of voiced speech sounds and is interrupted by voiceless sounds and speaking pauses. $F_0$ is inseparably connected with the underlying segmental string of an utterance, typically nuclear vowels, syllables, words, and phrases.

**Segment Duration.** Strictly speaking, this notion also includes the variation in the duration of a speech segment (phoneme, syllable, pauses) against the average duration for this segment in a given speech database (sometimes also called 'duration contour').

**Segment Intensity** which in an utterance can be associated with the 'envelope' of the speech signal.

In the scope of this thesis we concentrate on $F_0$ and duration, as these are the features commonly manipulated in text-to-speech systems. Syllable intensity, though it is one of the
output parameters of the integrated model will not be investigated systematically, nor will be other suprasegmental features such as voice quality.

As indicated above, the term ‘$F_0$ contour’ denotes the $F_0$ contour in its temporal relationship to meaningful units of speech (typically words) and the timing of intonational events will be described relative to these units.

For this reason all figures of $F_0$ contours displayed feature the locations of segment boundaries marked with vertical dotted lines as shown in Figure 2.1, where the key to the figures of $F_0$ contours henceforth employed is given. In Section 2.7 a more systematic definition of the relationship between $F_0$ contour and meaningful speech segments will be presented. The $F_0$ contour is consistently displayed in the log $F$ domain$^2$. From Section 4.3 on, this diagram will be augmented by a display of the duration contour in terms of a syllable-based $z$-score.

2.2 Speech Production

2.2.1 Fundamental Frequency ($F_0$)

![Fig 2.2: The muscles and cartilages of the larynx. The tension of the vocal cords is actively influenced by the vocalis muscle itself and passively by the cricothyroid muscle changing the relative position of the cricoid and thyroid cartilages. From Helfrich (1985).](image)

The source of voiced speech sounds is the oscillation of the vocal cords, a mucous membrane with muscular parts, the vocalis muscle, inside the larynx that separates the sub-glottal volume (lungs and trachea) from the vocal tract. The larynx is a cartilaginous structure suspended in

$^2$Throughout this work the term 'log' denotes the natural (base $e$) logarithm.
Figure 2.3: Muscles attached to the arytenoid cartilages controlling the degree of closure of the glottis (left: cricoarytenoid muscle (posterior), center: Lateral cricoarytenoid muscle (antisus), right: arytenoid muscle and lateral cricoarytenoid muscle. From Hobohm (1993).

Figure 2.4: Two degrees of freedom in the movement of the cricoid cartilage: Translation (left) and rotation (right) against the thyroid cartilage caused by activity of the cricothyroid muscle producing a change in $F_0$. From Zemlin (1968).

Figure 2.5: The shape of the glottal opening for a) respiration b) voicing and c) whisper. From Kahle (1984).
a mesh of antagonistic muscles which connect it with other parts of the skeleton and ensure its functionality independent of head or neck position.

The major parts of the larynx (see Figure 2.2) are the thyroid, cricoid and arytenoid cartilages and the epiglottis. The cartilages are connected by a number of intrinsic larynx muscles. The vocal cords are suspended horizontally between the cricoid and thyroid cartilages. The position of the pair-wise arytenoid cartilages at the rear end of the glottis determines its degree of closure.

The arytenoid cartilages are controlled by the cricoarytenoid muscle (posterior) which dilates and the lateral cricoarytenoid and arytenoid muscles which contract the glottis (Figure 2.3 from [Hob93]).

The frequency of the glottal oscillation (i.e. the laryngeal frequency $F_0$) is basically determined by the oscillating mass which is controlled by the degree of tension of the vocal cords [Hei85, p. 31]. The tension of the vocal cords can either be actively increased by activity of the vocalis muscle or passively by the cricothyroid muscle which modifies the relative position of the cricoid and thyroid cartilages and hence causes a change in length of the vocal cords. As shown in figure 2.4, the movement of the thyroid cartilage has two degrees of freedom: 1) Rotation around the cricothyroid joint 2) Translation of the thyroid cartilage against the cricoid cartilage. The former is ascribed to pars recta, the latter to pars obliqua of the cricothyroid muscle.

Shape and size of the glottal opening vary considerably. For light respiration and whispered speech the membranous part is closed and only the small part between the arytenoid cartilages remains open (Figure 2.5c). For stronger respiration also the membranous part is dilated, producing a rhombic opening (Figure 2.5a from [K+84]).

Before voice onset, the vocal folds are adducted, but they do not have to be fully closed to initiate phonation. The velocity of the airflow from the subglottal volume is increased at the place of the glottal obstruction. Due to the so-called ‘Bemoulli effect’ this results in a negative pressure at the medial edge of the glottis which are subsequently ‘sucked’ together. As long as the subglottal pressure is sufficient it bursts the glottal closure. This slightly reduces the subglottal pressure, causing the rim of the glottis to collapse and shortly close (Figure 2.5b). Subsequently, the glottis opens and closes again, and a quasi-periodic oscillation starts, an effect very similar to playing a comb.

The frequency of the glottal oscillation is controlled via auditory feedback and can be kept relatively constant even by untrained subjects\textsuperscript{3}.

### 2.2.2 Syllable Duration

The stream of air modulated by the glottis enters the vocal tract which functions as an acoustical filter enhancing or attenuating parts of the spectrum of the glottal waveform. The vocal tract (see Figure 2.6), consists of the pharynx and the oral and nasal cavities which are connected through the velar valve. The geometry and hence the transfer function of the vocal tract changes with the movements of the articulatory organs, i.e. the tongue, velum, jaw and lips. The characteristics of vowel sounds are mainly determined by the position of the tongue (low/high, front/back) and the presence or absence of lip rounding. Consonants are produced when the tongue or lips constrict the vocal tract, and differ as to the place of constriction (velar, palatal, alveolar, dental or labial) and the mode of articulation: Plosive (complete closure), fricative (close approximation), for instance.

Connected speech can be thought of as a fluid transition between states (or targets) characterizing the speech sounds we wish to produce. Hence, in the course of producing an utterance the speech organs move along spatial trajectories. As a consequence, sounds which we script by the same letter are not necessarily always articulated in the same way, but depend on the

---

\textsuperscript{3}Hanson [Han78] reports a standard deviation between 2 and 3 % for the fundamental period of a single tone.
neighbouring sounds. In German, for instance, a well-known example is the /x/ phoneme in words like ‘kochen’ and ‘риchen’, which is pronounced in a velar position after back vowels, such as [U] and [O] and in a palatal position after front vowels such as [i:] and [ɛ:]. The reason for these differences can be attributed to the minimization of articulatory effort and the underlying principle is called ‘coarticulation’.

Quasi-stable portions of the speech signal corresponding to so-called articulatory targets usually never last for more than between about 50 and 200 ms. This fact has led some researchers to suspect (c.f. [Ohm00], [Eng98]) that the segmental approach to speech is largely motivated by the textual representations of language which we use, and that it is rather the trajectories and transitions between steady states which characterize speech. In their view coarticulation is a mere construct of the ‘segmentalists’. Works by Local [Loc94] and others have shown that the accurate modeling of fine phonetic detail in a non-segmental approach to speech can largely improve the naturalness of synthetic speech.

An important property of speech is the interplay between the high intensity of vowels and the lower intensity of consonants which produces rhythmic patterns, easily identifiable by the amplitude modulations of the speech signal envelope as shown in Figure 2.7, top. The frequency of these patterns roughly corresponds to the natural frequency of the jaw movements of 6±1 Hz.

An infant exploring his/her speech organs usually starts off discovering the effect of opening and closing his/her mouth producing the well-known /ma-ma/ and /pa-pa/ patterns. This leads us immediately to the concept of speech syllables which basically consist of a vowel and possible preceding and/or following consonants. The consonants preceding the vowel are often referred to as the syllable ‘onset’, those following as the syllable ‘coda’. The part combining the so-called nuclear vowel and the coda is often referred to as the ‘rhyme’.

\footnote{Throughout this thesis SAMPA phonetic notation is employed, see page xii for a complete list of German SAMPA symbols.}

Figure 2.6: The vocal tract. Articulation involves the tongue, jaw and lips. From Hess (2000).
Although widely used and commonly accepted, the concept of syllables is not an accurately defined one. We do not want to replicate the underlying discussion here, but indicate some possible interpretations on an example and explain which notion of syllable is adopted in this thesis.

On the level of the speech signal (Figure 2.7), the acoustical level, syllables can be roughly delimited by the relative intensity minima between the vowels pertaining to consecutive syllables. As a consequence, syllable boundaries may well cut through individual speech sounds (indicated by parentheses).

On the phonological level, in every language, there exits a limited set of permitted syllable structures. [svE], as in [dasvEt@6] for instance, is not a legal syllable of German (except for certain loanwords), but in English or Swedish completely acceptable. In this thesis, we adopt the phonological syllabification. In figure, the ‘t’ in ‘Wetter’ is labeled separately and set in angled brackets in order to indicate that it belongs to both, the syllable [vEt] as well as the syllable [t@6]. The concept of two syllables sharing a common phone is called ambisyllabicity and throughout this thesis it is adopted for single inter-vowel consonants after a short vowel only. A stronger concept of ambisyllabicity is followed in [OLC99] where all intervocalic consonants which can form the legal coda or onset of a syllable are marked as ambisyllabic. Accordingly, syllable boundaries may overlap.

On the morphological level, the level defined by the smallest meaningful units, morphemes may define different boundaries than those of the phonological syllables, as indicated for the word ‘abschließend’ where the boundary is after the [s].

The duration of speech sounds is strongly influenced by the physiological constraints of the
vocal tract. Whereas most sounds, such as monophthong vowels and fricatives can be sustained for many seconds, there are limitations as to the highest speed at which they can still be uttered intelligibly. At a high speaking rate, certain phones tend to be reduced or disappear completely while some of their features may be inherited by adjacent phones [Koh99].

Hence, depending on speech rate, the number of syllables in a word may be reduced. Speech production research has shown that the syllable is an important temporal unit in the process of utterance planning.

2.3 Speech Perception

2.3.1 Perception of Intonation

Anybody may have experienced that in an environment where several speakers are talking at the same time, it is possible to selectively ‘tune in’ and follow a particular conversation. This phenomenon is known as the ‘cocktail party effect’ and alludes to the function of intonation as the ‘carrier wave’ of speech which works much like the carrier wave of a radio transmitter.

In the process of verbal communication, a major role of intonation is its prominence-lending function by which salient information is acoustically marked in the speech signal. It was shown that the degree of prominence of a syllable in an utterance can basically be increased by manipulating either $F_0$ or the duration of the respective syllable [FKN94]. This has two consequences:

1. The percepts of $F_0$ movements are easily confused with percepts of segments with higher energy (accounting for the inaccuracy of auditory transcription of intonation).

2. The speech-organizing function of intonation (in non-tone languages) can only partially be ascribed to its melodic properties, for instance, for the discrimination of questions from declarations.

Helfrich 1985 [Hel85] conducted perception experiments to examine the role of intonation as an intermediate step in the process of speech perception.

The first experiment (after Abrams & Bever [AB69]) deals with the localization of click sounds whose position in a test utterance was varied. The results show that the click is seldom identified at the point in the utterance where it actually occurs but the perceived location shifts to varying degrees depending on its distance from inter-clause boundaries which therefore must present perceptual cues for segmentation. Since, as Helfrich observes, the shapes of $F_0$ movements at inter-clause boundaries do not correspond to singularities or transients in the contour — as similar movements also occur within a clause — the significance of these shapes can only be explained as a result of evaluation of the partial pitch contour assigned to the clause. This evaluation requires a ‘storage’ of the pitch for a duration between 1 and 2 seconds. Helfrich suggests that the storage may well be connected to the association of $F_0$ movements with lexical elements.

In a second experiment (after Treisman [Tre64]) the subject is exposed to two parallel acoustic stimuli (two read utterances presented to the left and right ears spoken by different speakers). The subject is asked to repeat one of the utterances, the other one is declared irrelevant and to be ignored. The delay between the utterances is varied as well as the text underlying the utterances, which may be identical for both. After every trial the subject is asked if he thought the texts presented were equal or not.

The results of the second experiment suggest that the storage of $F_0$ for at least 2 s is possible without the listener identifying the lexical items in the utterance. This may facilitate the backward directed ‘correction’ of inadequate hypotheses as to the syntactic and semantic
contents of an utterance and thus generally enhance the intelligibility of speech in the process of speech perception.

Helfrich concludes that $F_0$ patterns serve as auditory units on the level of the sensory memory. They constitute transitory steps in the perception process towards higher-level (linguistic) representations.

A different kind of perceptual experiments was conducted by 't Hart [tH84] and the IPO group and later by d’Allessandro and Mertens [Md95]. Whereas the former proposed linear ‘copy contours’ which they claimed produced the same perceptual impression as the original contours, the latter examined which changes in $F_0$ caused a change in perception and determined thresholds. These approaches are discussed in Section 2.6 on intonation models.

2.3.2 Perception of Rhythm

Earlier descriptions of prosody have made the claim, that speech rhythm observes clock-like regularities. This concept, also termed ‘isochrony’ is based on the hypothesis that humans possess internal clocks which determine the pace of all biological processes, from the cellular metabolism to complex activities such as speech. One obvious pacemaker that can be thought of immediately is the heartbeat.

Following this concept, the notion of ‘stress-timing’, for instance, describes the principle that certain languages, among them English and German, observe a constant temporal distance between stressed syllables. Other languages, such as Japanese, are following temporal regularities on smaller level, the syllabic or mora level. The model of speech rhythm proposed by Dell and Nieminen [dDN99] attempts to account for this classification by positing that the observable speech rhythm is a result of an interaction between stress-timing and syllable timing which behave as a pair of coupled oscillators. The classification of languages as either stress or syllable timed would then be expressed by a coefficient $\tau$ determining the coupling strength between the two oscillators. Dell and Nieminen do not relate their model to the underlying physiological processes.

Instrumental investigations on the phenomenon of isochrony have shown a fairly high amount of temporal variation in rhythmic patterns considered as isochronous. Moreover, little proof has been found that speech behaves isochronously, neither on the syllabic nor on the stress group level [vS97b]. Even in fairly regular tasks, such as list-reading, little evidence was found that speakers utter words with less syllables considerably slower in order to observe an equidistance between subsequent stressed syllables [Pir98].

The fact that yet words with three, four or five syllables on the average exhibit shorter syllable durations than mono- or bisyllabic words can as well be attributed to the fact that the structure of the long word is more easily recognized than that of a short word and therefore individual syllables are simply less important [vS97b] and tend to be reduced.

More recent studies indicate that notions of stress- and syllable timing are rather influenced by the ratio of consonant and vowel portions in a language %V and the standard deviation of those portions $\Delta V$ and $\Delta C$, with ‘stress-timed’ languages exhibiting a smaller and more variable vowel portions. Perceptual studies with delexicalized speech proved that native speakers of a language L1 can consistently differentiate between languages L2 and L3 pertaining to different types [RN+99].

Other perceptual studies (for a summary, see [BB97]) suggest that observed timing regularities in speech are rather related to the perception of rhythm than manifest in the speech signal proper. In tapping experiments, the subjects usually produced a beat that preceded the click of the metronome. The delay between the beat and the click was longer for foot than for hand tapping and roughly corresponded to the difference in nerve conduction time from brain
2.4. MEASURING PROSODIC FEATURES

Whereas the $F_0$ contour $F_0(t)$ and intensity $I(t)$ ('envelope') can rather straightforwardly be extracted directly from the speech signal, measuring durations presupposes the segmentation of the speech signal into (phonetically or phonemically defined) portions. This segmentation can, for instance, be performed auditorily, by playing back small portions of the speech signal and delimiting the speech sounds in the oscillogram. Due to the inertia of the aural system, however, this kind of approach is relatively coarse. Visually delimiting speech sounds in the oscillogram proper yields better results. In addition, the short-time spectrum and formant trajectories can be taken into account. Current methods for database labeling usually comprise all methods mentioned so far and are facilitated by a set of rules defining segmentation criteria (see, for instance, [TKZ94]).

![Figure 2.8](image)

Figure 2.8: Transition between a vowel and a voiceless fricative. Voicing ceases about 15 ms after frication starts.

These criteria need to be formulated very carefully, as the transition between segments is often not exactly defined in the speech signal. If we have a look at the transition between a vowel and a voiceless fricative, for instance, voicing may cease long after frication starts (see Figure 2.8).

The estimation of a duration contour $D(t)$ requires the establishment of the relationship between the duration of a particular segment and the durations of segments of the same type in the entire database, the z-score, for instance: $z_i = (\text{dur}_i - \mu_i)/\sigma_i$. 

to hand vs. brain to foot. A tentative explanation for this observation is that the synchronicity between the two events (beat and click) is rather established on the perceptual level than on the production level.

The point to where the perception is anchored is also called ‘p-center’ and is roughly defined by the onset of the nuclear vowel. The unit defined by this kind of segmentation is also called inter-perceptual center group and is the basic time unit in the duration model proposed by Barbosa and Bailly [BB92] (see Section 2.6.2.7).

Pfitzinger [Pfi99] has shown that speech rate can be reliably assessed by subjects and that the perceived reading speed can be predicted by taking into account both the local syllable and phone rate.
Table 2.1: Information conveyed by prosody.

<table>
<thead>
<tr>
<th>modifying meaning</th>
<th>not modifying meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>linguistic (lexical, syntactic, semantic)</td>
<td>paralinguistic</td>
</tr>
<tr>
<td>sentence mode</td>
<td>speaker’s intention, attitude</td>
</tr>
<tr>
<td>discourse organization (focus)</td>
<td></td>
</tr>
<tr>
<td>segmentation (integration, delimitation)</td>
<td></td>
</tr>
<tr>
<td>disambiguation</td>
<td></td>
</tr>
</tbody>
</table>

2.5 Information Encoded in Prosodic Features

2.5.1 Functions of Prosody

Speaking of the functions of prosody, one has to take into account that speech is characterized by the cooccurrence of various features, like Fo, intensity, duration, voice quality etc., and neither one of these is the only correlate of the functions we assign to it.

Hébrard [Hé185] distinguishes between those functions of prosody which modify meaning and those which do not (see Table 2.1). The former could also be seen as the part of information which is consciously and intentionally provided by the speaker ([Fuj97]), the ‘message’, whereas the latter involuntarily accompanies it.5

The linguistic features concern the way a message is formally coded and organized into (prosodic) units of a certain language. They correspond to the ‘surface structure’ of the message on a still rather abstract level. The actual meaning of the message can often not be decoded without interpreting the underlying paralinguistic information.

The question “Are you tired?”, for instance, is simply a request for being supplied information on someone’s psychological and physiological condition. If it is asked with a concerned undertone then the message may be: “Come on, you’ve been working so hard, you have to get yourself some sleep!” With an ironical undertone, it may mean “You lazy guy, you’ve been sleeping all day and still you’re tired!”

An important ‘non-linguistic’ information encoded in prosodic features not listed here, is that speech is coming from a single, human speaker, and not a machine. This information could also be described as the degree of ‘naturalness’ which is attributed to an utterance by the listener.

Apart from the naturalness issue, the current work is mainly focussed on the linguistic features in Table 2.1, as the author is convinced that the formal characteristics of prosody need to be described appropriately, before an attempt can be made to examine paralinguistically influenced varieties. The issue of perceived naturalness, however, is always in play, especially when we evaluate synthetic speech.

Fujisaki [Fuj92] proposes a systematic model of the process by which information is coded in prosodic features of speech. He takes into account that, for instance, intonation and its acoustic correlate, the Fo contour, are the result of a complex multi-stage process, which is subject to certain constraints at each of its steps (see Figure 2.9 from [Fuj93]). Information from higher level processes (‘Input Information’) is coded into abstract units and structures of a particular language, here called ‘message planning’. The message planning is guided by the rules and

---

5Speakers, of course, are capable of disguising their dialect consciously or, on the contrary emphasize it, to mark their membership to a certain group. Actors are well capable of simulating various emotional conditions. This kind of voluntary control of otherwise non-linguistic features may rather belong to the paralinguistic features.
2.5. INFORMATION ENCODED IN PROSODIC FEATURES

Figure 2.9: Processes by which various kinds of information are coded in the segmental and suprasegmental features of speech [Fig95]. Higher level input information (linguistic, paralinguistic, non-linguistic) is transformed into speech sounds by a multi-stage process: 1) message planning, 2) utterance planning, 3) motor command generation and 4) speech sound production.

constraints which could also be called the ‘grammar’ of the language. In the next stage, an utterance is planned, taking into account the phrasing, accentuation and pausing principles of the particular language. On this step, paralinguistic and non-linguistic information first enter the production process, determining, for instance, the style and segmentation of the utterance.

The utterance plan leads to the generation of neuro-motor commands for controlling the speech production mechanism. The contents of information in the utterance is converted into acoustic correlates like accents, phrases and pauses. Both stages, command generation and the speech production mechanism are characterized by physiological and anatomical constraints (finite time constants, limited repertory of articulatory movements, limited range of $F_0$ etc.).

In the following an overview of terms employed in studies of prosody is given. These terms will be critically discussed which are employed with different meaning by different scholars.

In cases where terms simply represent synonyms for the same feature, synonyms are given and one term will be consistently employed for the scope of this study.

2.5.2 Word accent

Early observations of spoken German revealed that every word of German when uttered in isolation has at least one syllable which is most prominent. This led to the assumption that this syllable was connected to a higher intensity than the others and therefore was called a ‘stressed’ syllable bearing the ‘word accent’ or ‘lexical accent’.

Empirical studies, however, indicated that word accent syllables generally feature a co-occurrence (in descending order of importance) of a distinct $F_0$ movement, a duration longer than unaccented syllables and a peak of intensity [Rus91]. The contribution of these acoustic features may vary due to the syllable structure (short vowels $\rightarrow$ shorter syllable duration) or the absence of $F_0$ (whispered speech). There exist minimal pairs of words (mainly verbs) in German which are segmentally identical and only differ with respect to the location of the lexical accent. Fig. 2.10 shows the example ‘kunghen’ (to handle) vs. ‘umgehen’ (to avoid).

$^{6}$Unlike the conventional notions of ‘paralinguistic information’ [Lav93], Fujisaki defines it as the information which cannot be inferred from the written counterpart of an utterance, and which is deliberately added by the speaker.
The rules by which the accented syllable of a German word can be determined will not be discussed here. They largely depend on the origin of the word and there exists no complete formulation for all lexical items. In the case of simple native words of German, the first syllable of the stem, generally the penultimate of the word, carries the word accent [Koh77]. A number of function words (prepositions, articles, conjunctions) are regarded as being unaccented [SZ82, p. 44].

### 2.5.3 Accent, Stress, Pitch Accent and Stress Accent

In traditional linguistic literature, the terms `accent` and `stress` have often been employed synonymously; to describe the degree of prominence of one syllable in an utterance against the others [Phe81, p. 830]. As already explained, it was assumed for German that the prominence was caused by a higher intensity of the accented syllable. For this reason, German is still often classified as a ‘stress accent language’ as opposed to a ‘pitch accent language’ such as Japanese [HK95].

In contrast, recent empirical studies discussed the occurrence of ‘pitch accents’ in English, a language which also traditionally belonged to the group of stress accent languages [Pie80].

Gronnum [Grø90] employs the terms ‘stressed syllable’ and ‘sentence accent’, with the sentence accent being the stressed syllable which is most prominent in an utterance. She claims that both, inter alia, are connected to changes in $F_0$.

The author feels that the inconsistency in the use of the terms ‘accent’ and ‘stress’, at least as far as German is concerned, makes it necessary to reconsider them with regard to the results of our empirical phonetic studies.

As shown in Figure 2.10, the lexical accent of an isolated word of German is connected with $F_0$ movement, longer duration and intensity of the accented syllable.

The realization of word accent in disyllabic words of German and Japanese was compared by Mixdorff and Fujisaki [MF95b] and it was found that $F_0$ patterns of words in both languages may look rather similar. The most important difference between the languages is that syllable durations in Japanese are little affected by the location of the pitch accent whereas a shift of the word accent from the first to the second syllable in German words changes them significantly.
This leads us to the conclusion that in contrast to the so-called pitch accent languages, \( F_0 \) is not indispensable for marking the word accent location in German since duration and intensity shift with it.

So what is the function of \( F_0 \)? In figures 2.11 and 2.12 two utterances of the same sentence “Er kann damit umgehen.” — “He can handle it.” are displayed. In the first utterance, a simple statement is made: “You, know, he can handle the situation, no problem.” In the second, the auxiliary verb ‘kann’ — ‘can’ is being emphasized and the connotation could be something like: “My God, why do you underestimate him all the time? I told you, he can do it!” The \( F_0 \) pattern in Figure 2.11 on ‘umgehen’ resembles the one for the word in isolation (Figure 2.10), whereas it is rather flat in Figure 2.12. If the piece of speech pertaining to ‘umgehen’ in the latter utterance is listened to in isolation, it can still be identified as ‘umgehen’ against ‘umgehen’, because the ratio of duration and intensity of the syllables in the word remains the same. Returning to the original meaning of ‘stress’ this is exactly what can be found here: The word accent syllable is prominent because of its higher energy, it is ‘stressed’.

For this reason the term ‘accented’ will henceforth be reserved to those words of an utterance which feature a distinct \( F_0 \) movement on the word accent syllable. If the \( F_0 \) movement is absent, the word accent syllable is still stressed, but the word itself is ‘de-accented’. Hence in this definition for German, the terms ‘stress’ and ‘accent’ are not relational dichotomies but the former simply contains a subgroup of prosodic features of the latter:

<table>
<thead>
<tr>
<th>term</th>
<th>prosodic features</th>
</tr>
</thead>
<tbody>
<tr>
<td>accent</td>
<td>( F_0 ), duration, intensity</td>
</tr>
<tr>
<td>stress</td>
<td>duration, intensity</td>
</tr>
</tbody>
</table>

This view is shared by Shuijter [Sh95] for English who shows that stress and accent also differ in their vocal source features.

### 2.5.4 Accent Group, Prosodic Phrase, Sentence Accent

In utterances containing more than one word unacceptable words are prosodically linked to accentuated words. This either occurs proclitically, to a following accentuated word (article + noun, for instance) or enclitically, to a preceding accentuated word (verb + pronoun, for instance). This produces the smallest meaningful prosodic units, of which a longer utterance consists. These units are called ‘accent-groups’ [SZ82, p. 55] or ‘tone groups’ [Phe81, p. 856]. Generally accent groups are combined to form larger prosodic ‘phrases’ building a sentence.

In an utterance containing several accentable words, the word accent syllables are potential locations of prominence. The word accent syllable which is most prominent compared with the others in the utterance and/or marks the communicatively most important part, is called the ‘sentence accent’ or ‘core accent’ [SZ82, p. 48]. In Grønnum’s terminology [Gro90] the sentence accent necessarily is the acoustically most prominent. This definition, however, seems hardly applicable to German, since generally the last accent of an utterance, which is often acoustically less prominent, determines its meaning. This will be illustrated by the following example.

In Figures 2.13 and 2.14 two utterances of the sentence “Der Wagen war an der Wiese.” — “The car was on the lawn.” are displayed. Whereas in the first utterance information is given concerning the place where the car was found (last accent on ‘Wiese’ — ‘lawn’), in the second utterance emphasis is laid on the fact, that the car was found on the lawn, and not, for instance, the bicycle. In both cases, however, the word ‘Wagen’ — ‘car’ is the one with the largest \( F_0 \) movement and hence the acoustically most prominent.
Figure 2.11: Speech waveform (top) and $F_0$ contour (center) of the utterance “Er kann damit umgehen.” — “He can handle it.” ‘umgehen’ accented, neutral statement.

Figure 2.12: Speech waveform (top) and $F_0$ contour (center) of the utterance “Er kann damit umgehen.” — “He can handle it.” ‘umgehen’ de-accented, variety emphasizing the capability of a person by putting the sentence accent on the otherwise de-accented auxiliary verb ‘kann’.
2.5. INFORMATION ENCODED IN PROSODIC FEATURES

Figure 2.13: Example for the placement of the sentence accent: Speech waveform (top) and $F_0$ contour (center) of the utterance “Der Wagen war an der Wiese” — “The car was on the lawn.” Neutral statement about a vehicle and the location where it was found.

Figure 2.14: Example for the placement of the sentence accent: Speech waveform (top) and $F_0$ contour (center) of the utterance “Der Wagen war an der Wiese.” — “The car was on the lawn.” Utterance emphasizing the type of vehicle that was found on the lawn.
In neutral, unemotional speech, the position of the sentence accent is generally determined by the underlying syntactic structure. Some examples will be given, with the sentence accent syllable printed in bold type.

Peter hat geschrieben. — Peter has written.
Peter hat einen Brief geschrieben. — Peter has written a letter.
Peter hat einen Brief an seinen Vater geschrieben. — Peter has written a letter to his father.

Among others, Stock [SZ82] and Pheby [Phe81] have formulated rules for determining the sentence accent in German.

2.5.5 Focus

The term ‘focus’ [ABO89, p. 267ff.] is strongly related with the sentence accent. It describes the semantic concept, by which parts of speech (parts of a sentence, single words or even syllables) receive more prominence than others.

It also applies to the kind of isolated sentences, that were given as examples for the placement of the sentence accent in the preceding section. It is, for instance, shown by the fact that the sentence accent shifts when the sentence “Peter hat geschrieben.” is expanded by the object “einen Brief”, presumably because of the higher amount of information in the noun.

In continuous speech the focus may be determined by the context preceding an utterance. It may emphasize new information or contrast it against old information. This background information is not necessarily given explicitly, but may be tacit knowledge of the speakers or even common sense. The salient part-of-speech, which is also called the ‘focus domain’, is marked by placing the sentence accent on a word pertaining to it, the ‘focus exponent’ [Höh82]. Depending on the extent of the focus domain it is called a ‘narrow’ or ‘broad’ focus, the latter extending over more than a single content word.

One example of narrow focus was already displayed in Figure 2.14: “The car (not the bicycle) was on the lawn.”

Not every change of the focal condition, however, brings about a change of the location of the sentence accent. This will be illustrated on an example from Altmann [ABO89, p. 277].

A: Gulda spielt Geige.
B: Nein! Gulda spielt Klavier.

1) A: Gulda plays the violin.
   B: No! Gulda plays the piano.

A: Was ist denn hier los?
B: Gulda spielt Klavier.

3) A: What’s happening round here?
   B: Gulda’s playing the piano.

A: Was macht eigentlich Gulda?
B: Gulda spielt Klavier.

2) A: By the way, what’s Gulda doing?
   B: Gulda’s playing the piano.

Altmann and his co-workers found in their study, that a double contrast as shown by the following example is generally not marked intonationally: “Sie läßt die Nina das Leinen weben vs. ‘...den Baumwollstoff färben.’ ” — “She lets Nina weave the linen.” vs. “...dye the cotton fabric.”

Stock 1982 [SZ82] does not employ the term ‘focus’, but distinguishes between ‘neutral’ and ‘contrastive’ intonation. According to his concept, neutral intonation can be predicted by the syntactic relationship between the constituents of an utterance.
The examples of contrastive intonation he gives [p. 59] correspond to the narrow focus condition discussed above.

2.5.6 Sentence Mode

The term ‘sentence mode’ [Alt87] denotes syntactic structures (sentences) which can be assigned certain functions (statement, yes-/no-question etc.). The features characterizing the sentence mode belong to four groups: a) specific words (wh-words in wh-questions, for instance) b) word order (the position of the finite verb, for instance) c) morphological marking (imperfect subjunctive in optative sentences) d) intonational marking. Intonational marking is especially important in those cases where a distinction by other features is not possible, namely when two sentences have identical wording (see example Figures 2.15 and 2.16: “Leihen wir den Wagen!” vs. “Leihen wir den Wagen?” — “Let’s rent the car!” vs. “Do we rent the car?”) The example also shows that the F0 contours do not only differ at the tail of the utterances (falling in the exhortation and rising in the question), but change their course already at the sentence accent location on ‘leihen’.

2.5.7 Declination

The term ‘declination’ denotes the global downwards trend, which generally can be observed on F0 contours of utterances of declarations. It causes the offset value of F0 of the contour to generally lie below the onset value (see, for instance Figure 2.16).

In longer utterances, generally a reset or readjustment of the declination line occurs.

2.5.8 Location of F0 peaks

Kohler [Koh91] observed that the perceptual impression caused by an utterance, where the position of an F0 peak relative to the sentence accent syllable is gradually shifted, does not change continuously, but is subject to abrupt changes at certain points. Hence he attributes the location of F0 peaks a phonological significance, namely a categorical distinction. As one possible interpretation, he associates ‘early’ peaks with established facts, ‘medial peaks’ with new information and ‘late’ peaks receive a connotation of surprise (see Figure 2.17). Kohler employed re-synthesized stimuli of one-accent utterances where an F0 peak was subsequently shifted from the left to the right and conducted two kinds of experiments:

1. The subjects are exposed to the stimuli in ordered sequence and must decide when they perceive a sudden change in the melody of the sentence.
   → More than 60 % perceived a clear change at one point in the sequence.

2. The subjects are exposed to individual utterances and must relate these to one of a set of given contexts.
   → Clear difference between early and medial peak utterances, less marked between medial and late peaks.

The acoustic difference between early and medial peaks is that in the former the F0 contour falls all across the syllable nucleus, whereas in the latter, part of the rise occurs within the syllable nucleus. It is, however, generally disputed if speakers systematically modify the location of F0 peaks to convey a certain meaning [Möb93, p. 49].
Figure 2.15: Example for sentence mode ‘yes/no-question’: Speech waveform (top) and $F_0$ contour (center) of the utterance “Leihen wir den Wagen?” — “Shall we rent the car?”

Figure 2.16: Example for sentence mode ‘exhortation’: Speech waveform (top) and $F_0$ contour (center) of the utterance “Leihen wir den Wagen!” — “Let’s rent the car!”
2.5. INFORMATION ENCODED IN PROSODIC FEATURES

Figure 2.17: The three $F_0$ peak patterns proposed by Kohler [Koh91], ‘$V_{on}$’ marking the onset of the nuclear vowel. According to Kohler, ‘early’ peaks denote established facts, ‘medial’ peaks new information and ‘late’ peaks surprise, for instance.

2.5.9 Conclusions

Hawkins [Haw95] summarizes the systematic problems resulting from the multi-dimensionality\(^7\) of the features influencing the (single-dimensional) $F_0$ contour:

1. There do not exist definitions for discrete units of intonation
2. Acoustic correlates of linguistic units are typically complex and influence longer portions of the $F_0$ contour
3. They generally contribute to more than one linguistic unit
4. They are highly variable (addition by the author)

This will be discussed on the example of the utterance “Er kann ihm damit umgehen.” — “Doing so, he can cheat him.” (Figure 2.18). The $F_0$ movement on ‘umgehen’ — inter alia — has the following functions:

1. Marking of the word accent (distinction against ‘umgehen’)
2. Delimitation (marking finality)
3. Sentence accent (center of information → focus)
4. Marking of the sentence mode (falling contour → statement)

If one follows Kohler’s view, it can be stated that the $F_0$ peak is a medial one (falling/rising $F_0$ contour on the accented syllable) and possibly signals new information.

It must be noted that so far we have discussed linguistic functions mostly with respect to the $F_0$ contour, for the simple reason that $F_0$ is the strongest and most powerful correlate of these functions. In terms of the meaning of an utterance, segment durations play a somewhat inferior role. Nevertheless, it can be shown that stressed syllables on the average are longer than unstressed ones, and that syllable lengthening signals prosodic boundaries. In this respect,

\(^7\)linguistic, phonetic, physiological, pragmatic etc.
Figure 2.18: Speech waveform (top) and F0 contour (center) of the utterance “Er kann ihm damit umgehen.” — “Doing so he can cheat him.” An example for the co-occurrence of linguistic information in the F0 contour. The portion of the F0 contour on the finite verb ‘umgehen’ — inter alia — carries the following information: 1) lexical (contrast against ‘umgehen’), 2) segmentation (marking finality), 3) sentence accent (prominence by distinct F0 movement), 4) sentence mode (falling contour for statements).

Segment durations are important cues of prosodic structure, as will be shown in Section 4.4.

2.6 Models of Prosody

2.6.1 Introduction

During the last three decades, several descriptions (‘models’) of prosody were developed to define the relationship between the linguistic units underlying an utterance and the F0 contour and the other prosodic features. The approaches differ considerably depending on their theoretical background (linguistics, experimental phonetics) and their application (linguistic research, speech technology, language education). Most of the studies only cover parts of functions of intonation discussed so far. Some of the most important models of prosody will be discussed in the current section. In Section 2.7 the quantitative intonation model by the author for German, MFGI, is revisited which was expanded to accommodate syllable durations as well as intensity.

Figure 2.19 illustrates the role of prosodic models as a link between linguistic structures and their acoustic manifestations, for instance, the F0 contour. In principle, two general methods can be distinguished: One type of model (from the left to the right in the figure) deduces a phonological description from a linguistic structure (typically, the syntactic surface structure) specifying accent levels and phrasal boundaries, transforming these into an abstract description of the duration and F0 contours which is then by application of phonetic rules converted into the actual F0 contour. This type of model generates F0 contours from higher-level linguistic information and hence the method can be called “a generative approach”. The opposite approach

\[\text{In Figure 2.19, the three varieties of abstract representations of the F0 contour ‘sequence of tone switches’, ‘tone sequence’ and ‘linear stylization’ are referred to only as examples. There exist, of course, other (possibly better) representations.}\]
2.6. MODELS OF PROSODY

Generative Approach

Linguistic information

Phonologic representation

Abstract F0 contour

F0 contour

Analytical Approach

Figure 2.19: The role of prosodic models as links between linguistic structures and their acoustic manifestations in the duration and F0 contours of utterances. The generative approach aims at producing F0 contours from higher level information (top-down), whereas the analytical approach infers higher level information from the observed F0 contour by means of some kind of abstraction (bottom-up).

(right to left) aims at abstracting from the observable prosodic features by means of approximation techniques, yielding phonologically relevant basic elements. These are then used to infer linguistic units and structures. Since this approach is based on the analysis of the prosodic manifestations observed, either mathematically, graphically or auditorily, it will be called an “analytical approach”.

Most studies of German prosody, combine generative and analytical elements. All approaches, though with varying degrees, refer to the analysis of observed speech data generally elicited with regard to some limited linguistic or phonetic problem. Elicitation tasks were, for instance, focal placement or the realization of various sentence modes as in Altmann et al. [Alt84].

Not all of the resulting models, however, cover the complete sequence of intermediate steps between linguistic structure and prosodic features, and some of the models describe the connection between the two only either in left-to-right (starting from the linguistic description) or right-to-left manner (starting from the prosodic data).

We are now going to discuss current approaches for predicting duration and F0 contours while mostly confining ourselves to approaches which have actually been applied in speech synthesis. The subdivision between duration and F0 already indicates that most approaches treat these features strictly separately, with F0 contours aligned with respect to the segmental durations predicted by the duration model.

2.6.2 Duration Models

2.6.2.1 Introduction

We first give a general overview of duration models in speech synthesis and discuss some of the more influential approaches later on in more detail. Common methods basically differ in the following aspects:

**Durational Unit Predicted.** The temporal units predicted by most current systems are either the phone (phoneme), often referred to as ‘segment’, or the syllable. Since eventually phone durations are required for the acoustic synthesis, all syllable-based models include
some kind of mechanism for calculating segment duration from the superordinate syllable duration. Barbosa and Bailly’s model [BB92], instead of the syllable, treats perceptually motivated IPCGs as basic units. An IPCG starts with the onset of a nuclear vowel and ends at the onset of the following nuclear vowel. The so-called ‘non-segmental’ approaches to duration modeling [OLC99] attempt to model the temporal changes in acoustic features of the speech signal (voicing, formant trajectories, etc.) associated with segments.

**Predictor Factors.** Every model uses a particular vector of input features which are extracted on the linguistic and phonetic levels. Most commonly employed factors, inter alia, include (1) on the segmental level: the properties of the phone to be synthesized and its neighbouring phones, (2) on the syllabic level: the degree of accentuation and the position in a higher-level unit, such as the foot or accent group, (3) on the phrase level: the location of a segment with respect to a minor or major boundary and the position of the phrase in a sentence.

**The Prediction Method.** The algorithms used for calculating a numerical duration value from the vector of input features can be roughly divided into rule systems and statistical approaches. Riedi [Rie98], see Figure 2.20, further subdivides the statistical approaches into parametric and non-parametric regression models. Whereas the structure of a parametric regression model in terms of how it processes the input factors is determined a priori (c.f. sums-of-products, general linear models), non-parametric regression models are developed by unsupervised training and the model structure is determined automatically (multi-layer perceptrons, CARTs). The main difference between rule-based and statistical models is that a rule system can be built on relatively little speech data. The formulation of the rules, however, requires a high amount of expert knowledge and considerable optimization effort by trial-and-error. In contrast, statistical approaches are built from a large amount of labeled speech data. Once the data is available, however, the training process is relatively effortless. Furthermore, the importance of individual factors can be
easily assessed by the way the statistical models prioritize them. Recent works by Jokisch et al. [JP98] examine a combination of statistical approaches (neural networks) and rules.

**Pause Prediction.** Some current approaches incorporate the prediction of speech pauses as part of the model, others treat pauses strictly separately.

**Speech Rate.** Many current TTS systems produce different speech rates by linearly scaling the durations output by the duration model. As has been shown by Zellner-Keller [ZK98], however, certain properties of a duration model need to be modified for different reading speeds. As the speech rate not only affects the duration of individual segments, but also the overall prosodic structure of an utterance, this kind of modification needs to take place on an earlier step of processing when the phrasal structure of an utterance is determined.

The major challenge for a duration model is to account for unknown, unlearned data. As van Santen [vS93] states, the paradox of 'labeled sparsity' makes it almost impossible to create a database that contains all possible vectors of feature-value pairs of input factors of a model, yet the probability of encountering a very rare vector in a short utterance is very high, as there are very many rare combinations. The number of combinations grows with the log of speech segments included in a corpus.

### 2.6.2.2 Klatt (1979), Kohler (1988)

The most influential rule-based duration model is the one developed by Klatt [Kla79]. It is based on the results of various experimental studies of segment durations and formulated in the following two equations:

\[
dur := (dur_{inh} - dur_{min}) \times \frac{PRCNT}{100} + dur_{min} \quad (2.1)
\]

\[
PRCNT := \frac{PRCNT \times PRCNT1}{100} \quad (2.2)
\]

Starting from a phone inherent duration \(dur_{inh}\) and a phone minimum duration \(dur_{min}\) which are stored in a look-up table for all phone classes, the duration difference between the two is modified by recursive application of the rules. Application of all rules, altogether eleven, yields the final duration of a phone in a particular context. \(PRCNT1\) stands for the expansion or compression factor assigned to various predictor factors.

One rule, for instance, prescribes, that vowels or syllabic consonants preceding a pause are lengthened by a factor \(PRCNT1\) of 140, another rule halves \(dur_{min}\) for unstressed segments.

Although we do not give a complete account of the rules, it is worth-while summarizing which factors are taken into account and assigning them to the different levels in the prosodic hierarchy:

**Phrase level**
- Pause insertion at sentence-internal main clauses and comma boundaries
- Clause-final lengthening
- Non-phrase-final and non-word-final shortening

**Word level**
- Shortening of segments in polysyllabic words
- Non-initial consonant shortening
• Shortening of unstressed segments
• Lengthening of emphasized vowels

Syllable and segmental levels

• Influence of the post-vocalic consonant
• Shortening of segments in consonant clusters
• Lengthening of vowels or sonorant due to preceding plosive aspiration

Following Klatt, a similar rule system was developed for German [Koh88] and with some modifications for French [BS87]. The Kohler system was employed in the Dresden TTS system for performing the perceptual evaluation described in Chapter 3.

2.6.2.3 Campbell (1992)

The duration model developed in [Cam92] makes the claim (also called the ‘Elasticity Hypothesis’) that the duration of all segments in a syllable is related by a sole and single factor $z$ to the superordinate syllable duration. The elasticity of a segment is described by the probability density function of its durations within a given speech corpus. The segment duration is therefore formulated in terms of its mean and standard deviation (in the log$\text{dur}$ domain) as follows:

$$dur_i = \exp(\mu_i + z \times \sigma_i) \quad (2.3)$$

The factor $z$ is referred to as the z-score. The superordinate syllable duration is then given as the sum of the individual segment durations:

$$dur_{syl} = \sum_i dur_i \quad (2.4)$$

By successive application of the two equations, also referred to as the ‘repartition algorithm’, the duration of the individual segment is calculated. Campbell posits that the rhythmical organization is mainly determined by factors above the segmental level, and that once the syllable duration is determined, the segment duration can be derived directly.

The original version of the model calculates syllable durations by a multi-layer perceptron using the following input factors:

Phrase level

• Break index (depth of phrase boundary following a syllable) — four levels

Foot level

• Foot Type: headed or defective

Word level

• Function vs. content word: two classes
2.6. MODELS OF PROSODY

Syllable and segmental level

- Stress — rated perceptually: four levels
- Type of nuclear vowel: four classes
- Number of segments in syllable: seven levels

In a later version of his model [Cam93], Campbell adds the following factors:

- Position of the phrase in the utterance
- Position of the syllable in the word
- Number of syllables in foot
- Syllable rhyme open or closed

For accommodating phrase-final lengthening effects which affect the rhyme of a syllable rather than the onset, Equation 2.3 is modified to the following form:

\[ \text{dur}_i = \exp(\mu_i + z \times \alpha \times \sigma_i) \]  \hspace{1cm} (2.5)

where \( \alpha \) is set to .75 for phrase-final contexts and otherwise 1.0. Campbell reports a correlation of .89 between observed and predicted syllable durations.

In order to model more accurately the effects of the various input factors on the individual segments in a syllable, Campbell [Cam93] develops a method for calculating probability density functions for all segment classes, given the possible combinations of factor-value pairs. Since not all segments occur in the database in all possible contexts, Campbell clusters the segments in broad-classes as to manner and place of articulation. Instead of calculating the individual distributions, quantized shapes for the broad-classes are calculated and then rescaled to match the individual segment distributions. From these \( \mu \) and \( \sigma \) for a segment in a particular context can be derived. The two-level model improves the correlation between measured and observed data to .93.

2.6.2.4 van Santen (1993), Möbius (1996)

Van Santen [vS97b] claims that durations need to be modeled on the subsegmental level and that this can be performed by applying a repertory of templates describing the durational characteristics of a segment in a certain context by a time warping function.

Van Santen criticizes Campbell’s syllable-based approach for not accounting for the observation that external factors such as accentuation or pre-final lengthening may affect onset and rhyme of a syllable to a different degree. He also challenges the elasticity principle by the results of a study on segment-syllable independence which shows that syllables are consistently lengthened when they accommodate phones with longer inherent durations.

Van Santen argues that adjacent segments in a diphone are compressed or expanded following a monotonous ‘expansion’ profile and not in a step-wise manner, i.e. assigning one warping coefficient for the first and another one for the second as is tacitly assumed in diphone synthesis (see Figure 2.21 for possible expansion profiles, with the bottom-most being the step-wise approach commonly adopted in speech synthesis).

Van Santen proposes a ‘sums-of-products’ model which although it is a statistical model incorporates expert knowledge concerning which factors influence segment durations and how they interact. Segment durations are hence calculated as a function of a feature vector \( d \) by the following equation:
Figure 2.21: Parameter trajectories (F1 vs. F2) for long and short contexts, time warping functions and expansion profiles. The step-wise expansion is the one applied in conventional timing models. From [vS97b].
\[ \text{dur}(d) = \sum_{i=1}^{K} \prod_{j=1}^{I_i} S_{i,j}(d_j). \] (2.6)

\( K \) is the number of product terms in the model, \( I \) the set of indices of factors included in the \( i \)-th product term. Product terms may contain one single parameter. Hence the Klatt formula can be interpreted as a special case of a sums-of-products model with \( K=2 \):

\[ \text{dur}_\text{phone} = (\text{dur}_{\text{inh}} - \text{dur}_{\text{min}}) \times \text{PRCNT}/100 + \text{dur}_{\text{min}} = S_{\text{phone}} \times \prod_{j=1}^{I} S_j + S_{\text{phone}} \] (2.7)

The inherent and minimum durations are a property of the phone (\( S_{\text{phone}} \)), whereas \( \text{PRCNT} \) incorporates the multiplicative influences of other factors on segment duration (\( S_j \)).

Generally speaking, the sums-of-products model subsumes the observation that factors influencing duration are either additive or multiplicative, and that the direction of the influence of a factor is invariant, provided all other factors kept constant. The depth of a phrase boundary, for instance, is always positively correlated with the observed duration of a phone preceding the boundary. Once the structure of a sums-of-product model is developed in terms of input factors and interactions between these, the parameter estimation for the model is performed with the least-squares method [vS93].

Van Santen claims that one main advantage of the sums-of-products approach is the property that a relatively small amount of data is sufficient for the model to reach an asymptote of generalizability (a few hundred data points). In comparison CARTs require as many as 10,000 data points to converge.

Although there is no complete account of factors incorporated in van Santen’s model, the following ‘core factors’ are mentioned in [vSS97]:

**Phrase Level**
- Position of phrase in utterance

**Word Level**
- Position of word in phrase

**Syllable Level**
- Syllabic stress, word accent
- Position in the word

**Segment Level**
- Current, preceding and following phone identity (post-vocalic consonant, for instance)
- Place and manner of articulation
- Position in the syllable

Van Santen reports a correlation between predicted and observed segment durations over a set of 41,588 segments of .93 for English, and Möbius of .90 for German on a set of 24,240 segments. Figure 2.22 shows a tree of the segment classes distinguished in the model for German from [MvS96].
Chapter 2. Prosodic Features of Speech

Figure 2.22: Phone categories differentiated in a duration model for German. From [BMvS96].

2.6.2.5 Zellner-Keller (1998), Siebenhaar-Rölli (2001)

The merit of Zellner-Keller’s approach to predicting syllable duration is that the speech rate is explicitly taken into account [ZK98]. In her treatment of the segments in a syllable, Zellner-Keller largely follows Campbell’s approach with the difference that, for overcoming the data sparsity problem, segments are clustered depending on their durational properties ($\mu$ and $\sigma$ in the database), and not with respect to articulatory distinctions. Her statistical evaluations show that the speech rate has a vast influence on the way segments behave and hence can be clustered.

The factors incorporated in the model are the following:

**Phrase level**

- Break index (Depth of phrase boundary following a syllable)

**Word level**

- Function vs. content word: two classes
- Mono- or polysyllabic word

**Syllable and segmental level**

- Position in phrase and type of phrase-tense value
- Number of segments in syllable
- Durational properties of segments in a syllable: 158 values
- Schwa/non-schwa nuclear vowel
These factors are subsequently entered into general linear models created on the segmental, syllabic and the phrase levels. In other words, the prediction error remaining after modeling the segmental level is modeled by the syllabic model, and then the remaining error, especially the pre-final lengthening effects, is taken care of by the phrase model. Zellner-Keller reports a mean correlation of .85 between observed and predicted durations and argues that this value corresponds to typical inter-speaker correlations. Siebenhaar-Rölli [SZKK01] has recently applied Zellner-Keller’s approach to German.

2.6.2.6 Local, Ogden et al. (1994)

![Diagram](image)

Figure 2.23: Partial representation of the syllable ‘sit’ showing the properties of a syllable in the linguistic representation. From [OC99].

The non-segmental approach to modeling timing in Text-to-Speech pursued by Local and Ogden differs from the approaches documented so far, as it is not based on an appropriate ‘set of input factors’, but on a rich linguistic representation of a sentence to be synthesized.

This representation is created following a linguistically motivated prosodic theory of English describing rhythm as the way syllables arrange in a foot (see Figure 2.23 for a partial representation of the syllable ‘sit’). Syllables can either be weak or strong, with strong syllables forming the head of a foot. Depending on their structure syllables are light or heavy, the latter being syllables with long vowels and/or consonant clusters in the coda. Onset, nucleus and coda in a syllable are described by a set of distinctive features such as voiced/unvoiced. Depending on the respective syllabic weights in a di-syllabic foot, the rhythmical structure can either be short-long or equal-equal. Intervocalic consonants are treated as pertaining to both syllables, as long as the constitute legal syllable codas and onsets. Therefore in this description syllables may overlap to a varying extent, accounting for phenomena of coarticulation. A speech database is labeled in a tree-wise manner relating all terminal nodes — i.e. the segments — to the superordinate units. Syllabic durations are calculated using a CART model.

Although the database at present only contains a limited number of possible linguistic structures, perception experiments have shown that predicting syllable durations from a rich phonological representation significantly enhances the perceived quality of a speech synthesizer.
2.6.2.7 Barbosa and Bailly (1992)

Barbosa and Bailly’s duration model [BB92] differs from the other approaches discussed so far as it attempts to model durations with reference to the perception of speech rhythm. As a consequence, the durational unit which their model predicts is the inter-perceptual center group (IPCG). The IPCG is delimited by the onset of a nuclear vowel and the onset of the following vowel. Similar to Campbell’s approach, the model proceeds in two steps: First the duration of the IPCG is computed by a sequential network constrained by an internal clock (basically the speaking rate) and then the duration of the segments are accommodated within the duration of the IPCG. Pauses are explicitly modeled within the approach [BB97]. The structure of the model is indicated in Figure 2.24. Whereas the clock and the sentence modality is kept constant throughout the utterance, the ramps indicate the extent of a prosodic unit: the longer ramp for the sentence, and the shorter ones for the prosodic group. Two input cells are reserved for the current and the following prosodic marker, such as accents. The remaining input cells concern the segmental structure of the IPCG.

Hence the following information is incorporated in the model:
Table 2.2: Models of German intonation. **Abbreviations** IR: Intonation research, Tl: Teaching intonation, SS: Speech synthesis, PL: Prosodic labeling, PE: Perception experiments.

<table>
<thead>
<tr>
<th>author(year)</th>
<th>$F_0$ description</th>
<th>based on</th>
<th>applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isacenko (1964)</td>
<td>tone switches</td>
<td>pitch perception</td>
<td>PE</td>
</tr>
<tr>
<td>Bierwisch (1966)</td>
<td>tone levels</td>
<td>syntactic/surface structure</td>
<td>SS?</td>
</tr>
<tr>
<td>Phleby (1981)</td>
<td>tone syllables, -groups</td>
<td>syntactic/surface structure</td>
<td>Tl</td>
</tr>
<tr>
<td>Stock (1982)</td>
<td>tone switches, intonemes</td>
<td>syntactic/surface structure</td>
<td>Tl</td>
</tr>
<tr>
<td>Bannert (1983)</td>
<td>minima, offset</td>
<td>phonological description</td>
<td>SS?</td>
</tr>
<tr>
<td>Altmann (1989)</td>
<td>on-offsets, mins, max</td>
<td>sentence model</td>
<td>IR</td>
</tr>
<tr>
<td>Adriaens (1991)</td>
<td>copy-contours</td>
<td>pitch perception</td>
<td>PE</td>
</tr>
<tr>
<td>Kohler (1977, 1991)</td>
<td>peaks, valleys</td>
<td>experimental, linguistic</td>
<td>IR, SS</td>
</tr>
<tr>
<td>Möbius (1994)</td>
<td>Fujisaki model</td>
<td>fitting $F_0$ contours</td>
<td>SS</td>
</tr>
<tr>
<td>Möhler (1998)</td>
<td>Codebook Quantization</td>
<td>fitting $F_0$ contours</td>
<td>SS</td>
</tr>
</tbody>
</table>

Table 2.3: Models of intonation developed for languages other than German.

<table>
<thead>
<tr>
<th>author(year)</th>
<th>$F_0$ description</th>
<th>based on</th>
<th>applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>T’Hardt (1984)</td>
<td>copy-contours</td>
<td>pitch perception</td>
<td>PE</td>
</tr>
<tr>
<td>d’Allessandro (1995)</td>
<td>copy-contours</td>
<td>pitch perception</td>
<td>PE</td>
</tr>
<tr>
<td>Fujisaki (1982, 1984)</td>
<td>Fujisaki-model</td>
<td>fitting $F_0$ contours</td>
<td>SS, PE</td>
</tr>
<tr>
<td>Taylor (1994)</td>
<td>Rise-Fall-Continuation</td>
<td>fitting $F_0$ contours</td>
<td>SS</td>
</tr>
</tbody>
</table>

Phrase level
- Boundary marker
- Sentence mode

IPCG and segmental level
- Accent marker
- Number of consonants in IPCG
- Number of consonants in coda
- Nature of the vowel

2.6.3 Intonation Models

Table 2.2 gives an overview of models of intonation developed for German with a brief classification of their properties, theoretical background and applications. Some of these models originate from approaches for other languages, listed in Table 2.3. In the following sections the various approaches will be described in further detail. The works by Isacenko and Stock, which form the background of MFGI will be touched on in Section 2.7.

When critically discussing earlier models, the fact remains that the object they describe is the same: Features of German intonation. Hence, while pointing out formal differences between
the model, an attempt will be made to find the essentials (in terms of ‘knowledge of German intonation’) they have in common.

2.6.3.1 Bierwisch (1966)

![Syntax Tree](image)

Das ganze Unternehmen ist nutzlos gewesen

Bierwisch's description of German intonation [Bie66] stands in the tradition of generative grammar theory. Starting from the syntactic tree structure of a sentence (Figure 2.25 A), transformation rules are applied which yield accent levels and 'intonation units'\(^9\) (intonational phrases). Initially, the levels of boundaries to the left and right of a constituent in a sentence are determined by the depth of the node in the tree structure the constituent is connected to. Intonation units are formed by successive, rule-guided deletion of these boundaries. This cyclical rule which may eventually remove all boundaries in an utterance is parameterized with a factor \(p\) which is an abstract unit reflecting external factors like speech rate and speaking style.

Taking into account, accent levels, phrase boundaries and 'syntactic intonation markers' (SIM)\(^{10}\), re-writing rules produce a phonetic transcription of the \(F_0\) contour which is described by the relative tone levels of the syllables in the utterance and the direction of the pitch movements at accented syllables (Figure 2.25 B).

The system of intonational rules developed by Bierwisch shows a large amount of experience and observatory skill, and — at least for the examples he presents — yields plausible results. Bierwisch, however, does not give the phonetic evidence for their relevance and leaves the question unanswered, which acoustical correlates in the \(F_0\) contour correspond to the boundaries of the intonation units he proposes.

---

\(^9\) German: ‘Intonationseinheiten’.

\(^{10}\) SIMs include, for instance, markers for question-final rise (Q) and statement-final fall to a low sentence accent syllable (D) which are derived from the deep structure of a sentence. Bierwisch, however, gives only tentative explanations as to the taxonomy of the SIMs.
2.6. MODELS OF PROSODY

Figure 2.26: Top-down model as proposed by Bierwisch (1966). The deep structure of a message is transformed into the surface structure of an utterance and ‘syntactic intonation markers’ (SIM). By application of accent and intonation rules these are converted into the observable acoustic correlates of an utterance ($F_0$, intensity, duration and pausing) [Bie66].

Figure 2.26 shows the intonation model developed by Bierwisch. Bierwisch’s approach has been successfully employed in a Swiss-German TTS system [TR88].

2.6.3.2 Féry, Uhmann based on Pierrehumbert (1980)

Pierrehumbert’s approach [Pie80] describes $F_0$ contours as a sequence of high (H) and low (L) tones. These are associated with accented syllables and prosodic boundaries. The actual $F_0$ range between ‘high’ and ‘low’ tones is only locally defined and subject to rule-determined reductions, called ‘downstep’ or ‘catathesis’. Following autosegmental phonology, which Pierrehumbert applies to intonation, an utterance is built of a number of parallel hierarchical layers, ‘tiers’, which consist of sequences of phonological segments. The ‘tone tier’ corresponds to the phonological elements which make up the $F_0$ contour.

Figure 2.27 shows a prosodic tree structure for an example sentence constructed by Féry [Fér88] who applied the tone sequence approach to German. Figure 2.28 shows the respective $F_0$ contour and tones labeled by Féry. The largest intonational unit of German proposed by Féry is the ‘intonation phrase’, which as the author explains, corresponds to Pheby’s ‘tone group’. It
Die Fjorde in Norwegen sind unbeschreiblich schön.

Figure 2.27: Hierarchical levels of tone sequence model for German as proposed by Fény (1988). The utterance “Die Fjorde in Norwegen sind unbeschreiblich schön.” — “The fiords in Norway are indescribably beautiful.”, is divided into intonation phrases π which consist of two tacts τ each. The first tact contains all syllables before the nuclear accent of the intonation phrase and the second one the remaining ones. In the tone tier accented syllables and syllables at prosodic boundaries are assigned high (H) or low (L) tones producing a somewhat impressionistic description of the $F_0$ contour.

is characterized by its most prominent syllable, the nuclear syllable. By default, the intonation phrase is made of two tacts, the first comprising all syllables before the nuclear syllable and the second the remaining. Fény does not specify the criteria for determining the boundaries of intonation phrases, she only assumes that equally prominent accents in an utterance may constitute separate intonation phrases. The examples of two-phrase utterances in her data (as in Figure 2.28) are split at the boundary between subject phrase and predicate phrase. The tonal inventory proposed for German is specified in Table 2.4 after Uhm [Uhm88]. ‘*’ denotes a tone linked to a prominent accent syllable, ‘%’ denotes a tone linked to a syllable left of a prosodic boundary. Compared with English [BP86, p. 256], the pitch accent shapes $H + L^*$ and $L + H^*$ are missing.

The selection of tonal patterns for the description of a particular $F_0$ contour seems to be rather based on intuitive choice than on objective criteria. In the utterance “Die Fjorde in Norwegen sind unbeschreiblich schön.” — “The fiords in Norway are indescribably beautiful.”, (Figure 2.28), Fény assigns the accented syllable ‘Nor-’ the tone $L^*$, although it is followed by a rise of the $F_0$ contour and therefore, according to Table 2.4, should rather be assigned $L^* + H$. In Section 5.4.5 the G-ToBI representation developed at the IMS Stuttgart [May95] will be discussed and related to Fujisaki model commands.

2.6.3.3 Phelby (1981)

Phelby’s description of German intonation [Phe81] was incorporated into a recent comprehensive work on German grammar by Flämig [Flä91]. It follows Kohler’s separation of utterances into tone groups (corresponding to Kohler’s ‘intonational units’) tacts, syllables and phonemes. In
2.6. MODELS OF PROSODY

![Diagram of intonation](image)

Figure 2.28: Example of an utterance labeled according to the tone sequence approach for German by Fery (1988). The figure shows the $F_0$ contour of the utterance “Die Fjorde in Norwegen sind unbeschreiblich schön.” — “The fyrds in Norway are indescribably beautiful.” At the bottom accent levels are indicated by the number of ‘x’ symbols printed over the text of the sentence. ‘|’ symbols indicate the boundaries of intonation phrases, ‘∥’ the boundaries of tacts.

In contrast to Kohler, every tone group is characterized by one of three distinctive ‘tone patterns’ on its most prominent syllable, the ‘tone syllable’:
1) falling 2) rising 3) sustained. Each of these three patterns has a strong and weak variety. According to Pheby, in the so-called “unmarked case”, every tone group correspond to one syntactic clause. In the “marked case”, a tone group consists of more than one syntactic clauses.

Tone groups may be separated by ‘linguistic pauses’. In the following example tone syllables are printed in bold face, and boundaries of tone groups denoted by ‘/’:

```
// Wir wollen mit dem Zug fahren.// der weniger voll ist. //
“We want to take the train, it is less crowded.”
```

Pheby calls the correspondence between tone groups and syntactic clauses ‘congruence’. If it does not apply (the marked case) a modification in the meaning of the utterance occurs:

```
// Wir wollen mit dem Zug fahren, der weniger voll ist //
“We want to take the less crowded train.”
```

Pheby postulates a relationship of ‘gravity’ between the constituents of an utterance which determines the location of the sentence accent. He points out that in the unmarked case the sentence accent is generally placed on the constituent word that is closest to the tail of the utterance. Furthermore, the gravity of component parts of a sentence is not only determined by their position, but also by their syntactic properties. Both, gravity and position in the utterance often coincide ($A < B$ denoting ‘$A$ weaker than $B$’):

- subject - indirect object - direct object (order)
- subject $<$ indirect object $<$ direct object (relationship of gravity)

If the stronger component is missing or replaced by a pronoun the sentence accent shifts:
Table 2.4: Tonal elements of tone sequence approach as proposed by Uhmann 1988.

<table>
<thead>
<tr>
<th>Function types</th>
<th>Phonolog features</th>
<th>Phonetic Realization</th>
</tr>
</thead>
<tbody>
<tr>
<td>accent tone</td>
<td>H* + T</td>
<td>falling F₀ contour throughout accented syllable, affecting following unaccented syllables</td>
</tr>
<tr>
<td></td>
<td>T* + H</td>
<td>rising F₀ contour throughout accented syllable, affecting following unaccented syllables</td>
</tr>
<tr>
<td>Phonological correlate of focal features</td>
<td>H*</td>
<td>F₀ peak in the nucleus of accented syllable, no influence on following syllables</td>
</tr>
<tr>
<td></td>
<td>T*</td>
<td>F₀ valley in the nucleus of accented syllable, no influence on following syllables</td>
</tr>
<tr>
<td>boundary tone</td>
<td>T%</td>
<td>speaker-dependent F₀ minimum, (baseline) at phrase boundaries</td>
</tr>
<tr>
<td>Phonological correlate of phrasing rules</td>
<td>H%</td>
<td>F₀ values above T% and speaker-dependent ‘medium’ level</td>
</tr>
</tbody>
</table>

“Das Kind hat einem Mann eine Zeitung gegeben.”
subject  indir. object  dir. object  
“The child has given a newspaper to a man.”

“Das Kind hat sie einem Mann gegeben.”
subject (dir. object)  indir. object  
“The child has given it to a man.”

Selting [Sel93] criticizes that Phelby’s concept of tone groups as being too grammar-oriented and with little that is applicable to empirical analysis. The ‘linguistic pauses’ he proposes as prosodic cues to the boundaries of intonational units are rarely found in natural speech data.

2.6.3.4 Altmann, Batliner and Oppenrieder (1989)

The ultimate objective of Altmann and his co-workers 1989 [ABO89] was the definition of prototypes of German intonation. They addressed this problem by experiments examining the realization of sentence mode and focus in natural speech data. Two kinds of experiments were conducted:

1. Phonetic analysis of a great number of context-elicited natural utterances (F₀, duration, intensity)

2. Perception experiments with selected stimuli from 1): a) accent test (marking of prominent syllables), b) classification test (assignment of functional type, like statement, question etc.) and c) test of naturalness (checking for acceptability of a stimulus within a given context)

The speech material consisted of segmentally identical sentences, which can only be distinguished by intonation; corresponding to contrastive ‘minimal pairs’ defined by Altmann’s ‘sentence model’ [Alt84] (see Section 2.5.6). For the lack of an appropriate model of German intonation [ABO89, p. 6.7], the experimenters confined themselves to the extraction of on-and offset values and maxima and minima of F₀ in the phrases. In addition they estimated the duration and intensity of syntactic phrases. The authors dismissed the thought of applying the tone sequence approach by Pierrehumbert [Pie80] to the analysis of their data, because of its lack of objectivity.
2.6. MODELS OF PROSODY

Figure 2.29: Diagrams describing prototypical $F_0$ patterns produced by Altmann et al. [ABO89]. The y-axis gives $F_0$ in semi-tones relative to the base frequency of a speaker (the lowest frequency produced in the data). The black squares indicate averaged minimum and maximum values of $F_0$ for the second and third phrase of a three-phrase utterance. Duration is given in centiseconds. The average duration of the second and third phrases is indicated by the horizontal lines limited by white dots. The diagram display the following conditions: top left: declaration, focus on phrase 2, top right: question, focus on phrase 2, bottom left: declaration, focus on phrase 3, bottom right: question, focus on phrase 3.

By averaging over the extracted features the authors constructed diagrams for the prototypical patterns they proposed. Figure 2.29 shows some examples of these diagrams. The underlying utterances were produced from sentences like “Sie läßt die Nina das Leinen weben.” — “She's having Nina weave the linen.” with the possible focal conditions:

1. narrow focus on the object ‘Leinen’ (“FOKUS 2” in the figure)
2. narrow focus on the infinite verb ‘weben’ (“FOKUS 3” in the figure)
3. broad focus on object plus infinite verb ‘Leinen weben’
4. narrow focus on object ‘Leinen’ and narrow focus on the infinite verb ‘weben’

The diagrams in Figure 2.29 show the conditions 1) narrow focus on ‘Leinen’, declarations (top left), 2) narrow focus on ‘Leinen’, questions (top right), 3) narrow focus on ‘weben’, declarations (bottom left) and 4) narrow focus on ‘weben’, questions\(^\text{11}\).

The black squares denote minimum and maximum values of $F_0$ in the words ‘Leinen’ and ‘weben’. ‘n’ denotes the number of speakers over which the $F_0$ values are averaged. The

\(^{11}\text{A sample context for case 1: In einem Textilbetrieb läßt die Mutter der Angestellten ihre Tochter weben:" Mutter: “Was läßt die Meisterin meine Nina denn gerade weben?” Die Angestellte: “Sie läßt die Nina das Leinen weben.” — In a factory for textile goods. A mother asks an employee about the training progress of her daughter. Mother: “What’s the forewoman having Nina weave right now?” The employee: “She’s having Nina weave the linen.”}
values are normalized to a semi-tone scale relative to the speaker-individual minimum $F_0$. The horizontal lines delimited by dots denote the average durations of the items ‘Leinen’ and ‘weben’.

A summary of the main results of this study:

1. There exist main and secondary patterns for the same syntactic structure (‘Kerntypen’ and ‘Randtypen’).

2. Sentences with different contexts (narrow vs. broad focus) are not necessarily contrasted by intonation.

3. Questions are marked by high offset values of $F_0$, declarations by low offset values.

4. The perception of focal prominence may be influenced by two interacting psychological factors: Situational expectations (guided by the particular context of an utterance) and habitual expectations (‘the most important thing comes last’).12

It remains unclear, whether the prototypical patterns are any more than an impressionistic description of intonation. Quantifying extreme values of the $F_0$ contour, does not reflect, for instance, the changes in the contour and its relationship with the segments of an utterance and therefore cannot be used for synthesizing $F_0$. The author believes, however, that Altmann’s empirical approach based on intonational contrasts is quite attractive as it concentrates on cases where the intonational ‘load’ is maximal (see, for instance, the experiment described in Section 4.4 which adopts the principle.

2.6.3.5 Adriaens (1991), IPO

Figure 2.30: Example of a stylized $F_0$ contour from [tH84].

---

12Batliner calls the former factor ‘Einstellungseffekt’ and the latter ‘Deklinationseffekt’.
Adriaens' work on German intonation [Adr91] (cited after Möbius 1993 [Möb93, pp. 49 - 51]) is based on the studies by Cohen and 't Hart [tH84] for Dutch and English at IPO Eindhoven. The basic principle of the approach is the approximation of observed $F_0$ contours by piece-wise linear stylization, producing a 'copy contour' (see Figure 2.30 for an example). This is justified by the observation that not every minor change in the $F_0$ contour is perceptually relevant. The resulting simplified contour is said to be 'perceptually equivalent' to the original one, although the notion of 'equivalence' is not made clear. 't Hart reports that the stylized contour is not always perceptually identical to the original contour, but 'fully acceptable' as representative of 'normal intonation' [tH84]. Since the IPO model does not consider the linguistic units underlying the utterance, the question arises, how the term 'fully acceptable' should be interpreted, since an utterance and hence the corresponding $F_0$ contour can only be judged within a communicative context.

As a result of a heuristic investigation, Adriaens postulates a set of 12 standardized $F_0$ movements (seven rising and five falling) which differ as to the $F_0$ interval spanned, their position in the accented syllable and their duration. Seven of these basic patterns are connected to accented syllables and have prominence-marking functions, while the remaining five are only perceptually relevant. The patterns are superimposed over four different standardized declination lines. Adriaens does not investigate into the functional aspects of intonation.

It is obvious that the copy contours for a particular utterance largely depend on the phonetic judgment of the experimenter and hence lack objectivity.

2.6.3.6 d'Allesassandro & Mertens (1995)

D'Allesassandro and Mertens [Md95] developed a stylization approach for $F_0$ contours which takes into account the perceptual impression caused by a particular contour ('pitch'). Based on results from glissando research (the perception of pure tones of changing frequency) they state that a movement in the $F_0$ contour is only perceptible if it exceeds a certain threshold, the 'glissando threshold' (henceforth $G$). Also a change in the slope of the $F_0$ movement is perceptible only when it exceeds the 'differential glissando threshold' (henceforth $DG$). $F_0$ movements below these thresholds are perceived as static tones, which corresponds to a short-term integration of $F_0$. An utterance is segmented into syllable nuclei, and using an integration algorithm applying $G$ and $DG$ (see Figure 2.31) a linear stylization is produced for every syllable. The stylized contour is used to re-synthesize the utterance. The synthesized speech is then compared with the original. D'Allesassandro and Mertens found thresholds of $G = 0.32/\sqrt{T}$ and $DG = 20$ for continuous speech. They did not investigate into the linguistic content of the utterances used.

2.6.3.7 Bannert (1983)

Bannert's description of German intonation [Ban83] is based on the analysis of $F_0$ contours of context-delicited read speech. The corpus contained utterances of statements, yes-/no-questions and echo-questions produced by three speakers (see Figure 2.32 for an example, the regions of the nuclear vowels are marked by hatched rectangles and the onsets of vowels in accented syllables by vertical lines). The sentences were constructed by successively expanding the phrase "Die Männer" — "the men" by additional words: "Die Männer in der Menge" — "the men in the crowd", "Der Müller will die längeren Männer in der Menge nennen." — "The miller wants to name the taller men in the crowd." etc.\(^{13}\)

\(^{13}\)It is unclear, if the resulting all-voiced 'nonsense' sentences are useful for eliciting natural intonation, since the exclusive choice of the initial sounds [l] and [m] in the content words create a kind of stave-rhyme.
Figure 2.31: Block diagram of the pitch contour stylization algorithm [Md95]. PDA is the pitch determination algorithm. V/UV is the voicing decision.
2.6. MODELS OF PROSODY

Figure 2.32: Examples of $F_0$ patterns of statement (top), echo-question (center) and yes-/no-question (bottom) from [Ban83]: “Der lullende Müller von Lingen will die längeren Männer immer Lümmel nennen.” — “The lulling miller from Lingen always wants to call the taller men louds.” The vertical lines indicate the onset of the nuclear vowel in accented syllables.

Input to the model is a linguistically defined structure (a sentence) with binary prosodic features:

1. $\pm$ STRESS (syllable stress $\rightarrow$ lengthening of syllable)
2. $\pm$ LENGTH (of vowels)
3. $\pm$ ACCENT (rising $F_0$ movement on nuclear vowel of accented syllable)
4. $\pm$ TERMINAL (terminal: statement, non-terminal: question)
5. $\pm$ PHRASING (realization of phrase boundaries by means of $F_0$ movements, not specified)
6. $\pm$ CONTRAST (contrast expands $F_0$ range)
7. $\pm$ EMPHASIS (emphasis expands $F_0$ range)

The phonological component of the model uses these features to generate a) ‘basic temporal patterns’ (the speech segments) and b) ‘basic tonal patterns’ (qualitative $F_0$ contours). These are described by $F_0$ targets (high (H) and low (L) tones) which are further modified by the binary feature $\pm$ WIDE (wide for contrastive or emphatic conditions) and aligned with the segmental string. Non-final accented syllables are assigned L-tones, as well as last syllables in statements, whereas final accented syllables in statements and last syllables in questions are assigned H-tones.
The **intonation algorithm** of the model converts the sequence of H and L-tones into the desired $F_0$ contour. Bannert observed that if the $F_0$ range is expanded, this affects only the maxima in the contour while the local minima (the L-tones, temporally aligned with the consonant-vowel boundary in accented syllables, from where the $F_0$ movements start) remain relatively unaffected. He therefore uses these local minima to define the global declination line (German ‘Talline’ ‘valley line’) for his model. The characteristic $F_0$ movements on accented syllables are superimposed on the valley line. The intonation algorithm connects the $F_0$ targets determined by straight lines or cosine interpolation.

A third component of the model, the **modification component**, a kind of post-processor for the $F_0$ contour generated by the intonation algorithm (taking into account microprosodic influences and the like), is not elaborated by the author. Bannert concludes that his model requires perceptual evaluation.

### 2.6.3.8 Kohler (1977, 1991)

Earlier descriptions of German intonation by Kohler [Koh77] stand in the tradition of the English musical school represented by Halliday [Hal67]. These are inspired by musical descriptions of intonation which perceive $F_0$ contours as sequences of distinctive ‘tones’. Kohler subdivides utterances into ‘intonation units’, ‘tacts’, syllables and phonemes. A tact starts with an accented syllable, which is followed by unaccented syllables. An ‘intonation unit’ is a sequence of tacts that can be divided into an (optional) ‘pre-nucleus’ and the ‘nucleus’. The beginning of the nucleus is marked by the most prominent syllable in the utterance. For German, Kohler distinguishes between the following tones for the nucleus:

1. Tone 1: fall to a low level
2. Tone 2: rise to a high level
3. Tone 3: rise to a medium level
4. Tone 4: sustained medium level
5. Tone 5: fall followed by a rise to a medium level
6. Tone 6: rise followed by a fall to a low level

Möbius [Möb93, p. 43] calls into question whether this inventory of tones permits a complete description of German intonation patterns and if the tones represent intonational contrasts, a question left unanswered by Kohler.

Kohler 1991 [Koh91] introduces the Kiel intonation model (KIM), which generates an $F_0$ contour from a linguistic description of a sentence. Following Kohler’s earlier work, the $F_0$ contour corresponds to a sequence of peaks and valleys, which are aligned with the segmental string.

Like Bannert, Kohler defines a number of distinctive prosodic features which are either binary or graded and assigned to segmental units (especially nuclear vowels in a syllable, denoted <VOK>) or non-segmental units (phrasal boundaries, for instance).

Kohler distinguishes between two domains of German prosody: Stress and intonation. Any lexical item features a syllable bearing the lexical stress, which potentially becomes the location of sentence stress in an utterance. Sentence stress is acoustically marked by duration $<\pm\text{DSTRESS}>$ and $F_0$ movement $<\pm\text{FSTRESS}>$ assigned to $<\text{VOK}>$. The prominence of an item in an utterance is denoted by a digit between 0 and 9, 0 standing for unstressed words, 1 for secondary stress, 2–9 for primary stress with varying degrees of emphasis. Further markers denote de-accentuation $<\pm\text{DEACC}>$ (corresponds to ‘secondary stress’), emphasis $<\pm\text{EMPH}>$
('+' corresponds to prominence levels 3–9) and stress level \(< \pm \text{STRLEV} >\) which maps primary stresses 3 to 9 to stress levels 1 to 7. It remains unclear, why this kind of redundancy was introduced and why there exist exactly 10 different levels of prominence.

The stress features assigned to a particular utterance are determined using re-writing rules ('symbolic feature rules') which take into account the syntactic, semantic and pragmatic contents of the sentence.

![Diagram](image-url)

**Figure 2.33:** Environment of the Kiel intonation model [Koh91].

In the 'intonation domain' of the model, vowels with 'primary' and 'secondary' sentence stresses are assigned basic \(F_0\) patterns which are either 'valley' or 'peak'-shaped (\(< \pm \text{VALLEY}>\)). These contours are reminiscent of the tones discussed above and selected according to the features 'terminality' \(< \pm \text{TERMIN} >\), question/declaration \(< \pm \text{QUEST} >\), \(< \pm \text{EARLY} >\) and \(< \pm \text{LATE} >\) which describe the location of \(F_0\) peaks (early, medial and late) (in Section 2.5.8 the phonological meaning which Kohler assigns the location of \(F_0\) peaks was discussed).

Using the output of the intonation part of the model, parametric rules are applied to calculate \(F_0\) targets along the segmental string which also take into account the microprosodic influence of the speech sounds. Finally, cosine interpolation is applied to connect the \(F_0\) targets and produce a smoothed \(F_0\) contour. Figure 2.33 shows the environment of the Kiel intonation model [Koh91,
CHAPTER 2. PROSODIC FEATURES OF SPEECH

Table 2.5: Steps between linguistic information and $F_0$ contours covered by models of German intonation. LI: Linguistic information, PR: Phonological representation, AF: Abstracted $F_0$ contour, F0: $F_0$ contour. A: Analytical, G: Generative.

<table>
<thead>
<tr>
<th>author</th>
<th>approach</th>
<th>LI</th>
<th>PR</th>
<th>AF</th>
<th>F0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isaenko</td>
<td>G</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>Bierwisch</td>
<td>G</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>Pheny</td>
<td>G</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>Stock</td>
<td>G</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>Bannert</td>
<td>G</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>Féry, Uhmann</td>
<td>A/G</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>Altmann</td>
<td>A</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>Adriaens</td>
<td>A</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>KIM</td>
<td>G</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>Möbius</td>
<td>A/G</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
</tbody>
</table>

Table 2.6: Steps between linguistic information and $F_0$ contours covered by non-German models of intonation.

<table>
<thead>
<tr>
<th>author</th>
<th>approach</th>
<th>LI</th>
<th>PR</th>
<th>AF</th>
<th>F0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pierrehumbert</td>
<td>A/G</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>’t Hardt</td>
<td>A</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>d’Allessandro</td>
<td>A</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
<tr>
<td>Fujisaki</td>
<td>A/G</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
<td>ø</td>
</tr>
</tbody>
</table>

p. 340.

2.6.3.9 Discussion and Conclusions

Returning to Figure 2.19 it can be stated that none of the approaches discussed provides the means to both derive $F_0$ contours from linguistic structures and infer linguistic structures from $F_0$ contours. Tables 2.5 and 2.6 give a classification of the models and the range of relationships they describe. Most of the approaches have mainly generative character. For the sake of comparison, the original Fujisaki-model and its application to German by Möbius are included in the table.

Among all models discussed so far, undoubtedly, KIM is the most comprehensive descriptive system of German prosody. As Kohler explains, KIM is developed and further refined by a method of ‘interactive investigation’ using a TTS-system implementing the KIM rule system. This approach starts off with a definition of linguistic units and structures and explores their manifestation in the $F_0$ contour by evaluating speech produced by rule-guided synthesis. Because of its generative character, however, this method is less applicable to the analysis of natural $F_0$ data. It is useful though to perceptually validate hypotheses for new rules. The same is true for Bannert’s model.

Pierrehumbert’s approach holds the advantage of great simplification of the $F_0$ contour, but at the risk of little objectivity. The intuitive ‘ad hoc’ choice of tonal patterns does not solve the question as to the quantity of intonational events.

Altmann, Batliner and Oppenrieder produced quantitative descriptions of prototypical intona-
tion patterns of German, although they reduced the $F_0$ contour to so few characteristic values that the contour can hardly be regenerated.

The greatest benefit of Adriaens' model is the data reduction gained by a stylization method which preserves the general shape of the $F_0$ contour. There are, however, no criteria, by which a degree of stylization that preserves all important acoustic cues for linguistic information can be chosen. Mere 'acceptability' can hardly be a reliable measure. D’Allessandro and Mertens results show that movements in the $F_0$ contour have to exceed certain thresholds in order to be audible. It is, however, unclear if the perception of accents in an utterance can be directly compared with the perception of glissandi. The prominence of a certain syllable and hence the perceptibility of an $F_0$ movement is — inter alia — influenced by the duration and intensity of that particular syllable. Furthermore the perceptual impression caused by an utterance is certainly influenced by other factors, such as the expectation of the listener. Hence the linguistic content of an utterance must be taken into account in the analysis. Even if the stylization produced by the algorithm should be equivalent to perceived pitch, it is just another data reduction approach which does not yield prototypical patterns for generating $F_0$ contours.

The intriguing simplicity and elegance of Isaenko's approach can only be adequately judged if one listens to the stimuli he used in his experiments. Despite their technically poor quality they convinced the author of the present study of further examining the relevance of tone switches within a more sophisticated framework of German intonation. The importance of Stock's work is connected with his elaboration and functional specification of the intonational units defined by tone switches.

What do the approaches have in common?

1. They stress the importance of the course of the $F_0$ contour at accented syllables.
2. They emphasize the importance of the last accent in the utterance (the sentence accent) for signaling focal condition, terminality and sentence mode.
3. They observe de-accentuation after the sentence accent.

However, none of the models

1. solves the problem of how intonational domains in an utterance ('prosodic phrases') can be delimited (except for syntactic criteria) in the $F_0$ contour.
2. incorporates elements modelling the production process of $F_0$.
3. comes in the form of a mathematical formulation.

### 2.7 The quantitative intonation model MFGI

#### 2.7.1 Linguistic Background of MFGI

**Isaenko and Schädlich (1964)** The early work by Isaenko and Schädlich [IS64] who claimed that

"No theory (of intonation) can be based on events which are never repeated."

was one of the first consistent attempts to separate the syntactic functions of intonation from attitudinal or emotional ones which up to then had often been confused. It is based on perception experiments using synthesized stimuli with extremely simplified $F_0$ contours. These were designed to verify the hypothesis that the syntactic functions of German intonation can
be modeled using tone switches between two constant $F_0$ values connected to accented (‘ictic’) syllables and ‘pitch-interrupters’ at syntactic boundaries.

The stimuli were created by monotonizing natural utterances at two constant frequencies and splicing the corresponding tapes at the locations of the tone switches:

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Phrase</th>
<th>Vorbereitungen sind ge</th>
<th>alles ist beln</th>
<th>troffen</th>
<th>reit</th>
</tr>
</thead>
<tbody>
<tr>
<td>178.6 Hz</td>
<td>die Kinder vertrauen den Eltern</td>
<td>(question)</td>
<td>“The children trust the parents.”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>150 Hz</td>
<td>die Kinder vertrauen den Eltern</td>
<td>(unfinished)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Four types of tone switches are distinguished: 1) A fall before a low accented syllable  2) a fall after a high accented syllable  3) a rise before a high accented syllable and  4) a rise after a low accented syllable.

A) Sentence mode:

die Kinder vertrauen den Eltern

B) Focus:

die Kinder glauben dem Lehrer

C) Disambiguation:

ich weiß, daß der Mann im Auto schläft

D) Phrasing:

Peter arbeitet       //       verdient aber wenig

die Zeit schreiben    //       bringen Artikel und anwendungen

“Peter works, but earns little.”

“The journals carry articles and advertisements.”

Figure 2.34: Examples for syntactic functions modeled by tone switches from [IS64]. Underlining denotes low level of $F_0$, overlining high level of $F_0$. ‘//’ indicates pitch interrupters at phrase boundaries.

Figure 2.34 show examples where syntactic functions are modeled by using tone switches. The experiments showed a high consistency in the perceptual judgment in a large number of subjects ($N = 50$, ratings 64 – 82 %)

The experiments showed a high consistency in the perception of intended syntactic functions in a large number of subjects ($N = 50$, ratings 64 – 82 %).

Stock and Zacharias (1982) The tutorial on German sentence intonation by Stock and Zacharias [SZ82] further develops the concept of tone switches introduced by Isaenko.
It is based on impressionistic studies by Stock [Sto80] who defines phonologically distinctive elements of intonation which are called 'intonemes', in the tradition of the Russian intonation researcher Artemov [Art65]. Intonemes are characterized by the occurrence of a tone switch at an accented syllable.

(I) "Die Delegation fuhr nach Grunewald."

(II) "Wollt Ihr Pilze suchen?"

(N) "Sie gehen am Wald entlang, weil sie Pilze suchen wollen."

"The delegation went to Grunewald."

"Do you want to look for mushrooms?"

"They walk along the forest, because they want to look for mushrooms."

Figure 2.35: Examples of intonemes proposed by Stock (1982).

Depending on their communicative function, Stock distinguishes between the following classes of intonemes (see Figure 2.35 for examples).

1. **Information intoneme**: An intoneme with falling tone switch (utterance-final), signals the completeness of an utterance. Speaker’s main intention: Conveying a message.

2. **Contact intoneme**: An intoneme with rising intonation (utterance-final), marks questions which cannot be identified by the sentence structure. Speaker’s main intention: Establishing contact.

3. **Non-terminal intoneme**: An intoneme with rising tone switch to a medium level which is sustained after the accented syllable. It is usually found with non-final accents and signals the incompleteness of an utterance. Its use largely depends on speech rate and emotional condition.

Any intonation model for TTS requires information as to the appropriate accentuation and segmentation of an input text. In this respect, Stock’s work is extremely informative as it provides default accentuation rules (word accent, phrase and sentence accents), and rules for the prosodic segmentation of sentences into ‘accent groups’. These rules are applied in MFGI in order to develop hypotheses as to which words in a sentence of text will be accented, and where the locations of potential phrase boundaries are. Hence the most important of the rules will be documented here.

**Unaccetable Words** Articles, prepositions, conjunctions, auxiliary and modifying verbs, interrogative and relative adverbs, relative, personal, interrogative and reflexive pronouns are generally unaccetable.

**Core Accent in Attributive Groups (AG)** Nouns, adjectives or pronouns which are modified by an attribute, form 'attributive groups'. The communicatively most important accent (the core accent) is placed on the last accetable item in the AG. (Example: “ein gefeieter Künstler” — “a celebrated artist”).

Exceptions: VCGs whose complement is determined (identifiable, for instance, by the use of the direct article), VCGs consisting of an adverbial qualification followed by a past participle (Example: “wir werden das genannte Buch referieren” — “we will give a report on the book mentioned”).
Main Core Accent in Complex Utterances  In utterances containing several AGs or VCGs the main core accent (= the sentence accent) is placed on the last item bearing a core accent according to rules A 2 and A 3. See the following example with core accent syllables in bold type, main core accent underlined, VCGs delimited by [ ] and AGs by [:] : "[Der Vorsitzende des Ministerrats] [überreichte [den berufenen Professor]] [die Urkunden mit ihrer Ernennung]]" — "The chairman of the Council of Ministers presented the certificates of appointment to the appointed professors."

Formation of Accent Groups  Accent groups are the smallest meaningful prosodic units of an utterance. They are characterized by an accented item, which unaccented items are prosodically linked to. They are not interrupted by major syntactic boundaries. The linkage occurs either to a following accented item (proclisis) or to a preceding one (enclysis) (Examples, accent groups delimited by [ ] : "[Wir untersuchten es][mit der Lupe]." — "[We examined it][with the magnifying glass.").

Merging of Accent Groups  With increasing speaking rate, accent groups tend to merge. This occurs especially where a part-of-speech is subordinate to another. The boundary between theme and rhyme in an utterance is generally preserved.

Context-Dependent Shifting of the Core Accent  Depending on the context of an utterance, the core accent may shift to any component part of the sentence. This occurs, for instance, in cases where a contrast is expressed or special emphasis is laid.

2.7.2 The Fujisaki Model

The mathematical formulation used in MFGI for parametrizing $F_0$ contours is the well-known Fujisaki model [FH82], [FH84]. Figure 2.36 displays a block diagram of the model which has been shown to be capable of producing close approximations to a given contour from two kinds of input commands: phrase commands (impulses) and accent commands (stepwise functions). These are described by the following model parameters (henceforth referred to as ‘Fujisaki parameters’):

\[ Ap: \text{phrase command magnitude; } T0: \text{phrase command onset time; } \alpha: \text{time constant of phrase command; } Aa: \text{accent command amplitude; } T1: \text{accent command onset time; } T2: \text{accent command offset time; } \beta: \text{time constant of accent command.}\]

$Fb$, the ‘base frequency’, denotes the speaker-dependent asymptotic value of $F_0$ in the absence of accent commands.

The phrase component produced by the phrase commands accounts for the global shape of the $F_0$ contour and corresponds to the declination line of the contour. The accent commands determine the local shape of the $F_0$ contour, and are connected to word accents and boundary tones.

Phrase Component  $Gp(t)$ (equation 2.9) denotes the impulse response of the phrase control mechanism. The input signals to the phrase control mechanism are impulses, which are defined by their magnitude $Ap$ and their onset time $T0$. $\alpha$ denotes the time constant of the phrase control mechanism and is assumed as being constant within an utterance. The characteristic of the phrase component is displayed in Figure 2.37 for values of $Ap$ of .15, .3, .45 and .6 and $\alpha$ of 2.0/s. It can be seen that the steepness of the falling slope of the phrase component, which models the declination line of the $F_0$ contour, increases with increasing $Ap$. $Ap$ also determines the onset value of the $F_0$ contour relative to $Fb$, unless an accent command is present.
\[ \ln F_0(t) = \ln Fb + \sum_{i=1}^{I} A_{pi}Gp(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}[Ga(t - T_{1j}) - Ga(t - T_{2j})]. \quad (2.8) \]

\[ Gp(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & \text{for } t \geq 0, \\ 0, & \text{for } t < 0. \end{cases} \quad (2.9) \]

\[ Ga(t) = \begin{cases} \min \left[ 1 - (1 + \beta t) \exp(-\beta t), \gamma \right], & \text{for } t \geq 0, \\ 0, & \text{for } t < 0. \end{cases} \quad (2.10) \]

Figure 2.36: Block diagram of the Fujisaki model (Fujisaki and Hirose, 1984) and underlying mathematical formulations.

**Accent Component** $Ga(t)$ (equation 2.10) denotes the step-response of the accent control mechanism. The step-wise input signals to the accent control mechanism, the accent commands, are defined by their amplitude $Aa$, onset time $T1$ and offset time $T2$. $\beta$ denotes the time constant of the accent control mechanism and is assumed as being constant in an utterance. Figure 2.38 displays the response of the accent control mechanism to accent commands of 250 ms duration and $Aa$ of 1.0, .75, .5 and .25 ($\beta = 20.0/s$). The ceiling value $\gamma$ (typically set to .9) of the accent control mechanism ensures that the accent component reaches its maximum in finite time. Hence the change in $F_0$ is in proportion to $Aa$. Since the falling slope is modeled by the inverse function of the rising slope, it starts off steeper, accounting for the observation that falling $F_0$ movements are generally faster than rising, without the need to introduce a different time constant.

Figure 2.39 shows the response of the accent control mechanism to accent commands of $Aa$ of 1.0 and duration of 100, 150, 200 and 250 ms ($\beta = 20.0/s$). It can be seen that for durations 100 and 150 ms, due to finite $\beta$, the accent component does not rise to its full height and is slightly changed in shape. Hence, in these cases the proportionality between $Aa$ and the change in $F_0$ does not hold.

A main attraction of the Fujisaki model is the fact that it offers a physiological interpretation connecting $F_0$ movements with the dynamics of the larynx [Fuj88], as indicated in Figure 2.40.
This viewpoint is not shared by any of the currently used intonation models which mainly aim at breaking down a given \(F_0\) contour into a sequence of ‘shapes’ [Tay95, PKH95, MC98].\footnote{Taylor [Tay00, page 1711] posits that certain slowly rising contours of English cannot be modeled due to the falling shape of the phrase component. The author of this thesis has shown that by introduction of an additional component [Mix98, page 96] these phenomena, which are very rare in German, can be formally taken care of.} Furthermore, the Fujisaki model represents intonation hierarchically, i.e. on the accent and phrase levels, thus accounting for the fact, that there exist prosodic domains of different extension.

### 2.7.3 Components of MFGI

Following Isacenko and Stock, an \(F_0\) contour in German can be adequately described as a sequence of tone switches. These tone switches can be regarded as basic intonational elements. The term ‘intoneme’ proposed by Stock shall be adopted to classify those elements that feature tone switches on accented syllables\footnote{It is worthwhile noting that the standard accents used in German ToBI [May93], H*L and L*H, for instance, actually describe tone switches, not ‘singular’ tonal targets, but without assigning accents a quantity or even a function.}.

Further elements which are not necessarily connected to accented syllables, but are employed for marking syntactic boundaries, will be called ‘boundary tones’, a term proposed by Pierre-humbert [PB88].

Analogously with the term ‘phoneme’ on the segmental level, the term ‘intoneme’ describes intonational units that are quasi-discrete and denote phonological contrasts in a language.
Figure 2.38: Response of the accent control mechanism to accent commands of 250 ms duration and amplitude $A_a$ of 1.0, .75, .5 and .25 ($\beta=20$/s).

Figure 2.39: Response of the accent control mechanism to accent commands of amplitude $A_a$ of 1.0 and duration of 100, 150, 200 and 250 ms ($\beta=20$/s).
though the domain of an intoneme may cover a larger portion of the $F_0$ contour, its characteristic feature, the tone switch, can be seen as a discrete event. By means of the Fujisaki model intonemes can be described not only qualitatively, but quantitatively, namely by the timing and amplitude of accent commands to which they are connected.

In terms of Figure 2.19, intonemes and boundary tones belong to the Phonologic Representation, as they are distinct intonational elements which, at least for neutral utterances without narrow or contrastive focus, can be derived from the linguistic surface structure of a sentence. The phonetic realization of intonemes and boundary tones in terms of the observable $F_0$ contour makes use of the alignment information inherent in these intonational elements. Representing the Abstract $F_0$ contour level (Figure 2.19), Fujisaki model commands are aligned with the segmental string and parametrized using information yielded by the analysis of natural data (see [Mix98, p. 133 ff.] for the accent command amplitude difference between $N_\uparrow$- and $C_\uparrow$-intonemes, for instance). The $F_0$ contour follows automatically when the time-aligned commands are passed through the Fujisaki model.

As presented in the preceding section, there are three classes of intonemes: The information intoneme $I_\downarrow$, the contact intoneme $C_\uparrow$ and the non-terminal intoneme $N_\uparrow$.

Since the $I_\downarrow$-intoneme may also occur in utterances of questions, it does not stand in a statement/question opposition with the $C_\uparrow$-intoneme.

Two varieties of boundary tones were observed which are both connected to the question-final rise: The 'concatenating boundary tone' (henceforth $B_{cat}$) is produced by an accent command which concatenates with an accent command pertaining to a preceding $C_\uparrow$-intoneme (see Figure 2.44).

The 'non-concatenating boundary tone' (henceforth $B_{noncat}$) occurs where question mode is being marked locally (see Figure 2.43). The results of the perception experiment reported in [MF95a] suggested that the identification of an utterance as a question rather than an incomplete utterance is facilitated by the separation of the utterance-final accent command into one command with a lower amplitude pertaining to the $C_\uparrow$-intoneme, and one command with higher amplitude pertaining to the boundary tone $B_{cat}$.

Table 2.7 gives the perceptual, syntactic and pragmatic properties of intonemes and boundary tones.

A special case of the $I_\downarrow$-intoneme is the ('emphatic') intoneme $I_\downarrow(E)$ which corresponds to

---

16 The arrow indicates the direction of the tone switch.
Figure 2.41: Example of intoneme $I_i(E)$ in a statement.

Figure 2.42: Example of an $N\uparrow$-intoneme concatenating with an $I_i$-intoneme.

Figure 2.43: Example of an $I_i(E)$-intoneme in a question followed by a boundary tone $B_{nocat}$. 

("Sie haben den Wagen geliehen.")
the rise-fall intonation pattern found for narrowly focused items in statements (see, for instance, Figure 2.41.

It should be noted, that the repertory of intonemes presented here is motivated by the phonological distinctions that they represent, i.e. basically the linguistic function of signaling sentence mode. We cannot make any statements as to further functions (para-linguistic connotations, for instance) that will require a further subdivision of the intoneme types[17].

Further discussion is needed as to the question how the portions of the $F_0$ contour pertaining to a particular intoneme can be delimited. In an acoustic approach, for instance, an intoneme could be defined as starting with its characteristic tone switch and extending until the characteristic tone switch of the following accented syllable. In the present approach, however, a division of the $F_0$ contour into portions pertaining to meaningful units, i.e., words or groups of words is favored, as the location of accented syllables is highly dependent on constituency, i.e.

[17] Informal listening tests with resynthesized stimuli suggested that, for instance, an $I_↓$-intoneme late in the syllable might signal politeness.
the choice of words in an utterance and the location of their respective word accent syllables. Unlike other languages German has a vast variety of possible word accent locations for words with the same number of syllables.\footnote{Some examples for five-syllable words: 'Nebengewerbe', 'Gemüsehändler', 'Allerheilig', 'Modernisierung', 'Organisation' ('secondary occupation', 'greengrocer', 'All Saints' Day', 'modernization', 'organization').}

Hence the delimitation of intonemes is strongly influenced by the lexical and syntactic properties of a particular utterance. Some examples for the delimitation of intonemes from the data examined so far are given. Figures 2.42 to 2.44 show examples of utterances where all intonemes and boundary tones have been labeled. The domain in the $F_0$ contour pertaining to the respective intonational elements is indicated by the horizontal arrows.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{f0_contour.png}
\caption{Schematic model of $F_0$ contour with global declination which corresponds to the phrase component, and tone switches.}
\end{figure}

Analysis of natural $F_0$ contour showed that any utterance is invariably preceded by a phrase command, and further commands occur in utterance-medial positions mostly linked with syntactic boundaries, as we will see in Section 5.4.3. Hence, the term ‘prosodic phrase’ denotes the part of an utterance between two consecutive phrase commands. Therefore the number of prosodic phrases in an utterance corresponds to the number of phrase commands determined by the analysis with the Fujisaki model. In other words, the phrase component of the Fujisaki model corresponds to a ‘global declination component’ from which rising tone switches depart and to which falling tone switches return. This is illustrated by the schematic model of an $F_0$ contour shown in Figure 2.45.

If we go back to the original definition of tone switches by Isačenko, the question how intonemes and their underlying accent commands should be related with the segmental string appears relatively straightforward, i.e., their timing should be defined with respect to accented syllables and syllables bearing boundary tones. Since we can expect certain influences of the syllable structure on the precise timing of accent commands, this influence is studied in more detail in Section 4.2.

Also the phrase component can be readily anchored to a syllabic grid, as phrase commands precede phrase-initial syllables with a constant delay. Furthermore, as the results in [Mix98] show, the magnitude of the phrase commands is strongly correlated with the number of syllables in the preceding phrase.
Chapter 3

Perceptual Evaluation of MFGI

Abstract

This chapter documents the perception experiments performed for assessing the prosodic quality of the quantitative intonation model MFGI. Experiment designs applied are discussed in the first part of the chapter. The remainder of the chapter is dedicated to the series of perception experiments, laying emphasis on results indicating the need for an improved integrated model of prosody.
3.1 Introduction

The evaluation of prosodic quality in TTS systems is a rather new research area and there are very few commonly adopted procedures (see [Son99] for a detailed treatise).

For assessing speech quality in general we can distinguish between objective, machine-based and subjective methods employing human subjects. Works by Klaus and others [KFS97] have shown that measures which work very well for describing the quality of a telephone line (the signal-to-noise ratio, for instance) fail in the case of synthetic speech. Even noisy human speech is perceived as more natural than synthetic speech. The reason why objective measures fail in the case of speech synthesis is partly because natural speech possesses a certain redundancy that makes it intelligible even under adverse conditions and this property is not captured by automatic methods. Prosody is particularly difficult to assess because of its inherently communicative functions. On the other hand subjective evaluations are resource intensive and need careful design.

The main objectives of the experiment series documented in this chapter can be summarized as follows:

- The comparison of the MFGI implementation with other methods for generating $F_0$ contours in one and the same TTS system, DRESS [Hir95] developed at the Laboratory of Acoustics and Speech Communication, Dresden University of Technology, with respect to intelligibility, perception of intended accents and naturalness (Sections 3.3.4 to Section 3.3.7).

- The comparison of model-generated $F_0$ contours with contours copied from natural speech (Section 3.3.8).

- The comparison of DRESS as a whole system with other high-quality TTS systems for German (Sections 3.4.1 and 3.4.2).

The work described in this chapter was performed under DFG (Deutsche Forschungsgemeinschaft) research grant no. HO 1674, in collaboration with Prof. Dieter Mehrt (formerly Humboldt-University Berlin), who conducted most of the perception experiments [MM98].

3.2 Evaluation Designs Applied

A recently developed technique for comparing the prosodic quality of TTS systems is the PURR method proposed by Sonntag [Son99]. PURR stands for ‘Prosody Unveiling Restricted Representation’ and the approach basically converts speech signals into sinusoids varying in frequency and amplitude with the $F_0$ and amplitude of the speech signal. Voiceless portions of the speech signal are represented by pauses. The resulting signal is claimed to possess the prosodic features of the speech signal, but lacks the segmental information. Perceptual experiments have shown that the restricted representation can be used by subjects for linguistic tasks such as accent and sentence mood recognition, as well as naturalness ratings. Since the segments are missing, TTS systems of very different segmental quality can be compared with respect to their prosody.

Despite the advantages of using restricted representation methods, we decided to directly employ synthetic stimuli as they were produced by DRESS and other high-quality TTS systems. On the one hand, the first part of the experiments (documented in Sections 3.3.4 to 3.3.8) was conducted within the same framework (DRESS). All stimuli to be compared were generated using the same linguistic information, segment durations and were synthesized with the same diphone inventory. Hence we were sure to compare the quality of the intonation component.
On the other hand, we implicitly relied on the ‘training effect’ of the first experiments which gradually turned previously ‘untrained’ subjects (in terms of experience with synthetic speech) into ‘trained’ ones. The notion ‘skilled’ used when characterizing our subjects refers to subjects who by their educational background had phonetic knowledge.

As far as methods for conducting psycho-acoustic experiments are concerned, computer-based settings have the following advantages over the traditional method of tape playback and protocols on questionnaires:

- The subjects can listen to the acoustic stimuli as often as they like, which may facilitate the comparison of rather similar stimuli.
- Results are automatically logged in a format permitting an immediate computer-based evaluation.
- Speech samples can be easily randomized individually for each subject. This excludes a possible influence of the playback sequence on the results.
- Experimental designs can be easily modified.

A major disadvantage of computer-based experiments is the fact that depending on the number of computer systems available only few subjects can take part in the experiment in parallel which considerably raises the total duration of an experimental series. Therefore depending on external constraints, especially the availability of subjects, some of the experiments documented in this chapter were computer-based and others performed by playing back a tape.

We start the discussion with a description of the computer-based experiment designs.

### 3.2.1 Pair Comparison of Synthetic Stimuli

The subjects are supplied with headphones. After specifying their name, age and sex, a short description of the experiment design is displayed. Then the subjects are presented with an individually randomized sequence of pairs consisting of two examples of the same sentence. They are allowed to play back each of the examples as often as they like by clicking the playback buttons on the GUI (see Figure 3.1).

In each trial, the subjects have to listen to the stimulus at least once by clicking the playback button with a built-in track-ball. They have to respond by choosing between the options ‘A sounds more natural than B’, ‘B sounds more natural than A’ or ‘both sound equally natural/unnatural’, by clicking one of the decision buttons. In the right upper part the number of remaining trials is displayed as a progress indicator. Whenever a reply-button is clicked the program checks whether each sample has been played back at least once. Otherwise an error message is displayed. If both samples have been played back at least once, the selected choice is displayed again and subjects asked for confirmation. In the case of handling errors or doubts concerning the correctness of the decision, the subjects can return to the last pair of stimuli, play them again and correct their choice. After every 10 trials a jingle is played enforcing a short break. PSYACEX logs subjects’ choices and stores them to an ASCII-File. After completion of the experiment, protocols of all subjects are summarized in a table which then can be further processed by statistics software.

The program PSYACEX was used in the following experiments:

- Perceived Naturalness in a pair-comparison, Section 3.3.7
- Comparison of fully synthetic and reference examples, Section 3.3.8
- Pair-comparison of isolated Sentences in cross-system comparisons, Section 3.4.1.
3.2. EVALUATION DESIGNS APPLIED

Figure 3.1: Graphical user interface of program PSYACEX.

Figure 3.2: Graphical user interface of program RANKING.
3.2.2 Ranking TTS systems

The program TTS-RANKING was specially developed for the system comparison (Section 3.4.2). Subjects' goal is the establishment of a ranking based on the prosodic quality of several TTS systems compared. Different from the pair-comparison with PSYACEX, subjects have samples from all TTS systems accessible in one GUI and are requested to establish a ranking in a more or less exploratory fashion. They are asked to assign points to each system on a scale between 0 and 10, and are requested to give 0 points to the least acceptable and 10 to the best one. This condition is imposed in order to force subjects to make use of the whole range of the scale.

After specifying their name, age and sex, the subjects are presented with a GUI equipped with playback buttons for each TTS system (Figure 3.2). Subjects are allowed to play back each of the examples over headphone as often as they like. To the right of the playback buttons edit fields are located where the subjects are requested to enter the number of points they assigned to the respective systems. In an additional control field radio buttons are arranged for selecting a playback mode. Subjects can either listen to a complete text passage in one row or to sentences individually.

Playback can be interrupted at any time by clicking the button ‘Wiedergabe anhalten’. After subjects have listened to each approach at least once and assigned points according to the ranking they perceived, they click the button ‘Fertig’ done and hence terminate the experiment. The result of the subject’s judgement and all playback actions are stored in a protocol file.

3.3 $F_0$ Control within DRESS

As discussed in the preceding section, we deemed a comparison of complete TTS systems useful only in the second place, since it requires listeners to ignore, for instance, that systems differ with respect to

- segmental quality (which can be due to the synthesis approach, consider formant against PSOLA synthesis),
- degree of pleasantness of the synthetic voice,
- degree of correctness of accent placement and phrase segmentation,

and only concentrate on the prosodic quality. For this reason the preliminary experiments (Sections 3.3.4 to 3.3.8) were conducted within the framework of the TU Dresden TTS system (DRESS).

Table 3.1 gives an overview of all perception experiments discussed in this section.
Table 3.1: Overview of all perception experiments in this Chapter. ‘NN mod.’ denotes a modified version of the neural net trained on a larger corpus, ‘nat. dur.’ stands for ‘natural durations’.

<table>
<thead>
<tr>
<th>Section</th>
<th>1st Preliminary Experiment</th>
<th>2nd Preliminary Experiment</th>
<th>Main Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.3.4</td>
<td>3.3.5</td>
<td>3.3.7</td>
</tr>
<tr>
<td>Comparison</td>
<td>within DRESS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criterion</td>
<td>intelligibility</td>
<td>naturalness (grading)</td>
<td>naturalness</td>
</tr>
<tr>
<td></td>
<td>naturalness</td>
<td>accent perception</td>
<td>naturalness</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>naturalness</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(pair comparison)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(system comparison)</td>
</tr>
<tr>
<td>Approaches Compared</td>
<td>* MFGI</td>
<td>MFGI</td>
<td>DRESS/MFGI</td>
</tr>
<tr>
<td></td>
<td>* NN I</td>
<td>MFGI</td>
<td>DRESS/NN mod.</td>
</tr>
<tr>
<td></td>
<td>* NN II*</td>
<td>MFGI</td>
<td>alien system A</td>
</tr>
<tr>
<td></td>
<td>* Hirschfeld</td>
<td>MFGI</td>
<td>alien system B</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>Difference between NN I and II</td>
<td>MFGI</td>
</tr>
<tr>
<td></td>
<td>Amount of training data I &gt; II</td>
<td>MFGI</td>
<td>nat. speech</td>
</tr>
<tr>
<td>Source</td>
<td>tape</td>
<td>tape</td>
<td>tape</td>
</tr>
<tr>
<td>Speech Data</td>
<td>72 sentences</td>
<td>72 sentences</td>
<td>72 sentences (84 pairs)</td>
</tr>
<tr>
<td>Judgement</td>
<td>single</td>
<td>single</td>
<td>single</td>
</tr>
<tr>
<td>Number Subjects</td>
<td>10</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>Skilled/Unskilled</td>
<td>6/4</td>
<td>6/4</td>
<td>6/4</td>
</tr>
<tr>
<td>Male/Female</td>
<td>-</td>
<td>-</td>
<td>5-5/7-5</td>
</tr>
<tr>
<td>Result (Ranking of Systems)</td>
<td>intelligibility for all systems close to 100 %</td>
<td>1. MFGI</td>
<td>1. copy contours/nat. dur.</td>
</tr>
<tr>
<td></td>
<td>3. NN I</td>
<td>2. Hirschfeld</td>
<td>3. DRESS/NN mod.</td>
</tr>
<tr>
<td></td>
<td>4. NN II</td>
<td>3. NN I</td>
<td>4. alien system C</td>
</tr>
<tr>
<td></td>
<td>5. Hirschfeld</td>
<td>4. NN mod.</td>
<td>6. alien system A</td>
</tr>
<tr>
<td></td>
<td>6. NN mod.</td>
<td>5. Hirschfeld</td>
<td>7. alien system B</td>
</tr>
<tr>
<td></td>
<td>7. Hirschfeld</td>
<td>6. NN mod.</td>
<td>8. alien system A</td>
</tr>
</tbody>
</table>
3.3.1 The Experimental TTS system

Besides MFGI, two other approaches for intonation control which had been developed at TU Dresden were embedded into the Dresden TTS system DRESS for conducting the preliminary experiments. One of these was the rule-based approach by Hirschfeld [Hir96], which belonged to the original system, and the other one was an approach based on a neural network [JP97], henceforth short ‘NN’. Of the latter two types I and II were used in the first experiments which differed as to the amount of training data which was larger in the case of type I.

Figure 3.3 shows a block diagram of the TTS system which was used for creating the synthetic speech samples. The function of all components will be briefly discussed.

**Pre-processing** The input text is segmented into meaningful chunks, typically sentences, parts of sentences or title lines. In rare cases (when punctuation marks are missing, for instance), a phrase is terminated before certain function words (‘and’, ‘or’). Within a single chunk, henceforth called a ‘phrase’, the words of the text are marked, numbers divided into categories and abbreviations and special characters processed. Number formats such as integers, year, date or time are being pre-processed in order to facilitate their pronunciation in the following step, the Grapheme-to-Phoneme Conversion. Abbreviations are replaced by their full forms or marked for citation. The output of this step is a formatted text.

**Grapheme-to-Phoneme Conversion** For each word, each number and abbreviation the pronunciation is determined in the form of a SAMPA-based phonetic transcript. Each word is assigned word accent location and number of syllables. This information is extracted from annotations in the phonetic transcription, if the word can be found in the dictionary, or by applying a grammar. In addition, the word type is supplied if it is found in the dictionary.

**Phonetic Module** In this step, the phonetic transcription between consecutive words is being post-processed. This is necessary for applying ‘Liaison’-Rules, for instance, which may modify the ending of a word depending on the beginning of the next word. Besides, the accentuation information for the entire phrase is being post-processed. In English, there are cases when the accent location and pronunciation depends on the word type (noun ‘a process’ vs. verb ‘to process’) Furthermore, ‘long’ phrases are segmented in this module and accent rules applied.
3.3. $F_0$ CONTROL WITHIN DRESS

**Duration Control** For each sound in a phrase the duration is computed using a Klatt-based rule system (see Section 2.6.2.2) based on inherent and minimum duration, and the distinctive features of a particular phoneme, such as voicing or frication. The selected reading speed is taken into account by linearly scaling the output durations of the model.

**Intonation Control** In this step the $F_0$ contour is being calculated and output in terms of three $F_0$ values per sound at 0, 33 and 66 % sound duration with each of the three $F_0$ approaches tested.

**TD-PSOLA-Synthesis** Every sound is concatenated by overlapping two diphones. The diphone units most feasible for producing a given sound sequence is determined over a table mapping legal sound sequences on diphone units available. Based on this information a speech waveform is produced. The diphones are read from the sound inventory and modified concerning their durations and $F_0$. The resulting concatenated speech signal is then output.

For the experiments discussed in Sections 3.3.4 to 3.3.7 only the intonation control was modified (MF Shape, Hirschfeld, NN), all other components remained unchanged.

### 3.3.2 The Corpus Used

Appendix B contains a complete listing of sentences in the corpus used for the various perception experiments at the sentence-level (Sections 3.3.4 to 3.4.1). It consists of three parts containing examples of statement, non-terminal and question intonation. The corpus had been compiled by Dieter Mehnert at Humboldt University Berlin and was selected because of its coverage of different sentence types [Meh85]. Recordings by trained speakers of all sentences were available which we considered useful for reference purposes.

### 3.3.3 Subjects

A total number of 53 subjects was available for the perception experiments, most of whom were students of Humboldt University Berlin. 21 of these students were considered as phonetically skilled, had already taken part in similar experiments and were familiar with synthetic speech. The remaining 32 subjects did not have any experience with TTS systems.

Groups of subjects (skilled/uns skilled, male/female) were recruited according to subjects' availability and the importance of the particular experiment (small or large groups of listeners).

### 3.3.4 Intelligibility

In the first experiment the intelligibility of the synthetic speech samples was tested in order to examine and document that the synthesis quality with all approaches was high enough to justify the evaluation of prosodic details. In line with our expectations, all four approaches reached a sentence intelligibility between 92 and 100 % (see Table 3.2 for details).

This experiment, as well as the following experiments in Sections 3.3.5 and 3.3.6 were conducted by playing back a tape of randomized speech examples and using special questionnaires.

Intentionally every example was presented only once. The subjects were given sufficient time between examples for writing down the sentence they had heard.

### 3.3.5 Perceived Naturalness (Grading)

The second experiment which was performed using the same examples concerned the perceived naturalness. The subjects were requested to rate each example on a four-point scale (1 - excellent
Table 3.2: Total result on intelligibility.

<table>
<thead>
<tr>
<th>approach</th>
<th>mean value [%]</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFGI</td>
<td>100.0</td>
<td>0.00</td>
</tr>
<tr>
<td>Hirschfeld</td>
<td>99.2</td>
<td>2.53</td>
</tr>
<tr>
<td>NN II</td>
<td>99.2</td>
<td>2.53</td>
</tr>
<tr>
<td>NN I</td>
<td>96.8</td>
<td>4.13</td>
</tr>
</tbody>
</table>

Table 3.3: Results of naturalness judgement per subject.

<table>
<thead>
<tr>
<th>approach</th>
<th>subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MD</td>
</tr>
<tr>
<td>MFGI</td>
<td>1.25</td>
</tr>
<tr>
<td>Hirschfeld</td>
<td>2.42</td>
</tr>
<tr>
<td>NN I</td>
<td>2.91</td>
</tr>
<tr>
<td>NN II</td>
<td>3.17</td>
</tr>
</tbody>
</table>

to 4 - poor). Every example was presented only once in order to encourage a spontaneous reaction.

Table 3.3 and Table 3.4 summarize the result of the naturalness judgement. MFGI is unanimously rated best, the distance between MFGI and the ‘least acceptable’ approach (NN II), however, amounts to only one point, which means that the latter approach is still perceived as ‘fair’. This indicates that the approaches compared cluster rather closely on the naturalness scale. We partly attributed this outcome to the influence of the segmental quality - which was the same for all approaches - on the perceived naturalness, as the grading approach only allowed for a relatively coarse judgement, given that the subjects listened to each stimulus only once. For this reason, we decided to apply a pair comparison paradigm in the following experiments implying naturalness ratings.

3.3.6 Perception of Intended Accents

This part of the preliminary experiments concerned the perception of intended accents in the synthetic speech examples. Subjects were supplied with the text of the sentences to be presented on a protocol sheet. Each example was played back twice with a short intermediate pause. Subjects were asked to mark the locations in the text where they believed they had perceived an

Table 3.4: Mean result of naturalness judgement. Grades: excellent = 1, good = 2, fair = 3, poor = 4.

<table>
<thead>
<tr>
<th>approach</th>
<th>mean value</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFGI</td>
<td>1.69</td>
<td>0.36</td>
</tr>
<tr>
<td>Hirschfeld</td>
<td>2.21</td>
<td>0.33</td>
</tr>
<tr>
<td>NN I</td>
<td>2.42</td>
<td>0.39</td>
</tr>
<tr>
<td>NN II</td>
<td>2.65</td>
<td>0.37</td>
</tr>
</tbody>
</table>
Figure 3.4: Ratio of perceived and intended accents. Of the NN approaches only the better type I is displayed.

Table 3.5: Results of intended accent perception per subject.

<table>
<thead>
<tr>
<th>approach</th>
<th>subject</th>
<th>MD</th>
<th>RE</th>
<th>DI</th>
<th>OL</th>
<th>HL</th>
<th>AW</th>
<th>DE</th>
<th>RH</th>
<th>OL</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFGI</td>
<td></td>
<td>92.6</td>
<td>91.8</td>
<td>81.7</td>
<td>88.1</td>
<td>85.9</td>
<td>89.5</td>
<td>86.9</td>
<td>90.3</td>
<td>80.7</td>
<td>81.9</td>
</tr>
<tr>
<td>Hirschfeld</td>
<td></td>
<td>66.6</td>
<td>64.8</td>
<td>61.7</td>
<td>62.0</td>
<td>66.2</td>
<td>88.0</td>
<td>61.2</td>
<td>83.1</td>
<td>62.7</td>
<td>63.2</td>
</tr>
<tr>
<td>NN I</td>
<td></td>
<td>59.9</td>
<td>61.7</td>
<td>46.4</td>
<td>53.4</td>
<td>64.9</td>
<td>67.5</td>
<td>45.6</td>
<td>54.8</td>
<td>56.5</td>
<td>61.0</td>
</tr>
<tr>
<td>NN II</td>
<td></td>
<td>52.1</td>
<td>46.7</td>
<td>50.5</td>
<td>49.0</td>
<td>44.9</td>
<td>51.8</td>
<td>45.4</td>
<td>51.8</td>
<td>51.8</td>
<td>52.3</td>
</tr>
</tbody>
</table>

Table 3.6: Perception of intended accents.

<table>
<thead>
<tr>
<th>approach</th>
<th>mean value [%]</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFGI</td>
<td>86.9</td>
<td>4.33</td>
</tr>
<tr>
<td>Hirschfeld</td>
<td>67.9</td>
<td>9.52</td>
</tr>
<tr>
<td>NN I</td>
<td>57.1</td>
<td>7.29</td>
</tr>
<tr>
<td>NN II</td>
<td>49.6</td>
<td>2.93</td>
</tr>
</tbody>
</table>
Table 3.7: Approach-wise comparison, rows correspond to approach A, columns to approach B. The cells in the upper table denote the frequency of vote ‘approach A preferred over approach B’, the cells in the lower table the frequency of vote ‘approach A equally (un)natural as approach B’.

<table>
<thead>
<tr>
<th></th>
<th>MFGI</th>
<th>Hirschfeld</th>
<th>NN I</th>
<th>NN II</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFGI</td>
<td>-</td>
<td>257</td>
<td>304</td>
<td>398</td>
</tr>
<tr>
<td>Hirschfeld</td>
<td>88</td>
<td>-</td>
<td>225</td>
<td>392</td>
</tr>
<tr>
<td>NN I</td>
<td>123</td>
<td>185</td>
<td>-</td>
<td>347</td>
</tr>
<tr>
<td>NN II</td>
<td>62</td>
<td>76</td>
<td>126</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MFGI</th>
<th>Hirschfeld</th>
<th>NN I</th>
<th>NN II</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFGI</td>
<td>-</td>
<td>258</td>
<td>184</td>
<td>165</td>
</tr>
<tr>
<td>Hirschfeld</td>
<td>258</td>
<td>-</td>
<td>201</td>
<td>141</td>
</tr>
<tr>
<td>NN I</td>
<td>258</td>
<td>201</td>
<td>-</td>
<td>142</td>
</tr>
<tr>
<td>NN II</td>
<td>165</td>
<td>141</td>
<td>142</td>
<td>-</td>
</tr>
</tbody>
</table>

accent. For the evaluation the percentage of correctly heard intended accents, i.e. the accents output by the phonetic module, was calculated. Figure 3.4 displays a bar chart of percentages of intended and perceived depending on the $F_0$ approach.

Table 3.5 and Table 3.6 show clearly, that most subjects perceived over 85% of intended accents in the speech samples produced with MFGI. On the average, the Hirschfeld approach rates about 20% lower, but still ranks above the best NN approach.

3.3.7 Perceived Naturalness in a Pair-Comparison

The first series of preliminary experiments was concluded with a computer-based pair-wise naturalness comparison using PSYACEx, involving the four approaches and the following 14 sentences:

1. Heinz spielt.
2. Hören Sie.
3. Schwer ist das.
5. Das Gespräch zeigte die Gegensätze und die gemeinsamen Züge unserer Auffassungen.
7. Du mußt alle kennen, wenn Du sie beurteilen willst.
8. Wir fragen ihn gern um Rat, denn er weiß meistens eine Lösung.
10. Sobald der Dirigent den Taktstock hob, trat tiefe Stille ein.
11. Wir stimmen dem Vorschlag zu, weil Ihr uns überzeugt habt.
12. Er arbeitete unermüdlich, um die Höhe seiner Kunst zu erreichen.
13. Wo liegt das Buch?
14. Wo sich Kieselsäure bildet?
3.3. $F_0$ CONTROL WITHIN DRESS

Table 3.8: Approach-wise comparison, rows correspond to approach A, columns to approach B. The cells display the percentage of point assignment to approach A.

<table>
<thead>
<tr>
<th></th>
<th>MFGI</th>
<th>Hirschfeld</th>
<th>NN I</th>
<th>NN II</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFGI</td>
<td>-</td>
<td>64</td>
<td>65</td>
<td>77</td>
</tr>
<tr>
<td>Hirschfeld</td>
<td>36</td>
<td>-</td>
<td>53</td>
<td>76</td>
</tr>
<tr>
<td>NN I</td>
<td>35</td>
<td>47</td>
<td>-</td>
<td>68</td>
</tr>
<tr>
<td>NN II</td>
<td>23</td>
<td>24</td>
<td>32</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.9: Naturalness score depending on the approach (scale 0-6) and Thurstone-scaled values with MFGI as the reference.

<table>
<thead>
<tr>
<th>approach</th>
<th>mean value</th>
<th>standard deviation</th>
<th>scaled values</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFGI</td>
<td>4.62</td>
<td>0.43</td>
<td>0.00</td>
</tr>
<tr>
<td>Hirschfeld</td>
<td>3.30</td>
<td>0.81</td>
<td>0.38</td>
</tr>
<tr>
<td>NN I</td>
<td>3.00</td>
<td>1.20</td>
<td>0.53</td>
</tr>
<tr>
<td>NN II</td>
<td>1.58</td>
<td>1.25</td>
<td>0.72</td>
</tr>
</tbody>
</table>

The overall naturalness score for a particular approach is calculated as the number of points this approach acquired in all pair-comparisons averaged over all subjects. In every pair-comparison one point was assigned to the superior sample, and none to the inferior one. If both were assessed as equally natural, each of them received half a point. Since each system was compared with three others twice for each sentence, the maximum score is 6.

Table 3.7, top, lists the frequency with which system A (rows) was preferred over system B (columns) over all sentences and subjects, and, bottom, the frequency of cases in which both approaches were judged equally (un)natural. Comparison of the two tables reveals that subjects rather often resorted to the 'equal' vote. Table 3.8 summarizes the results in Table 3.7 in terms of percentages of points shared between the approaches.

Table 3.9 displays the total result which confirms the ranking established in the grading experiment. The standard deviation for the NN-based stimuli are considerably higher than for the rule-based approaches. In order to calculate the perceptual distance between the four approaches, Thurstone scaling according to [MS00] was applied to the results of the paired comparison. The result of the scaling is displayed in the right-most column of Table 3.9 with reference to MFGI.

3.3.8 Fully Synthetic against Reference Examples

The experiment in this section was designed in order to judge the relevance of results yielded in the preceding experiments by comparing the naturalness of synthetic $F_0$ contours and natural $F_0$ contours as a reference.

A direct comparison with natural speech samples was excluded at this stage, since this would have implied the modification in parallel of several features of the speech signal, such as segmental quality and segmental durations.
3.3.8.1 Speech Material Used

For this experiment the following sentences were selected from the corpus:

1. Bereitwillig gab er Auskunft.
2. Aller Anfang ist schwer.
5. Üben und immer wieder üben ist beim Erlernen jeder fremden Sprache notwendig.
6. Das Gespräch zeigte die Gegensätze und die gemeinsamen Züge unserer Auffassungen.
7. Wenn wir die Maschine anschalten, beginnt der Motor zu surren.
8. Es regnete soviel, daß der Fluß über die Ufer trat.

The reference examples were created by copying the $F_0$ contour from a natural utterance onto a synthetic example of the same sentence (Figure 3.5, top). In addition, stimuli were produced in which, apart from copying the $F_0$ contour, the segment durations were adjusted to those measured in the natural speech sample (Figure 3.5, center).

In order to determine a copy contour, segment boundaries were initially marked in the natural utterance (Figure 3.5, bottom). The $F_0$ contour was extracted in intervals of 10 ms, and then the timing of the $F_0$ samples was related to the duration of the speech segment in which they occurred.

The resulting segment-related $F_0$ values were then used to generate the synthetic sample. For compatibility reasons, the natural samples were uttered by the same speaker who had produced the diphone inventory used in the PSOLA synthesis of DRESS.

The NN-based approach used in this experiment had been trained on a larger speech corpus than the one used in Section 3.3.7 [JP98].

Figure 3.6 shows examples of speech samples produced with the full synthesis approaches. The comparison of synthetic and natural contours shows that the latter exhibit a significantly wider $F_0$ range and stronger microprosodic variations.

20 subjects were presented the resulting stimuli in sets of pairs in randomized order, using a tape organized according to the following pattern:

<table>
<thead>
<tr>
<th>sample A</th>
<th>2.5 s pause</th>
<th>sample B</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 s pause</td>
<td>sample A</td>
<td>2.5 s pause</td>
</tr>
<tr>
<td>sample B</td>
<td>10 s pause</td>
<td></td>
</tr>
</tbody>
</table>

Like in the preceding experiment (Section 3.3.7), the subjects' task was to either judge the naturalness of sample A as superior or inferior, or judge both samples as equally good or bad.

3.3.8.2 Results

Table 3.10 displays the sentence-wise results of the naturalness judgement. In every pair-comparison one point was assigned to the superior sample, and none to the inferior one. If both were assessed as equally natural, each of them received half a point. Since each approach was compared with four others for each sentence, the maximum score is 4. It becomes clear that
3.3. $F_0$ CONTROL WITHIN DRESS

Figure 3.5: Reference examples, for each example the upper panel displays the $F_0$ contour and the lower one the speech waveform and the phone boundaries (from top to bottom): Synthesis with copied $F_0$ contour, synthesis with copied $F_0$ contour and natural segment durations, original speech signal.
Figure 3.6: Full synthesis examples, for each example the upper panel displays the $F_0$ contour and the lower one the speech waveform and the phone boundaries (from top to bottom): MFG1, Hirschfeld model, and modified NN.
Table 3.10: Sentence-wise comparison. Abbreviations: MD: copy contour and natural durations, MA: copy contour only, MF: MFGI, NN: neural network, HI: Hirschfeld.

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Satz</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>5.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bereitwillig gab er Auskunft.</td>
<td>MD (2.63)</td>
<td>MF (2.50)</td>
<td>NN (2.25)</td>
<td>MA (1.60)</td>
<td>HI (1.02)</td>
</tr>
<tr>
<td>2</td>
<td>Aller Anfang ist schwer</td>
<td>MD (2.75)</td>
<td>MA (2.05)</td>
<td>MF (2.03)</td>
<td>HI (1.98)</td>
<td>NN (0.95)</td>
</tr>
<tr>
<td>3</td>
<td>Die Begründung ist stichhaltig</td>
<td>NN (2.25)</td>
<td>MF (2.23)</td>
<td>MD (2.08)</td>
<td>MA (1.93)</td>
<td>HI (1.53)</td>
</tr>
<tr>
<td>4</td>
<td>In dem Lehrbuch sind viele Hinweise...</td>
<td>MD (2.95)</td>
<td>MA (2.42)</td>
<td>MF (2.00)</td>
<td>NN (1.48)</td>
<td>HI (1.15)</td>
</tr>
<tr>
<td>5</td>
<td>Üben und immer wieder üben ist beim...</td>
<td>MF (2.93)</td>
<td>MD (2.30)</td>
<td>HI (2.03)</td>
<td>MA (1.73)</td>
<td>NN (1.02)</td>
</tr>
<tr>
<td>6</td>
<td>Das Gespräch zeige die Gegensätze...</td>
<td>MF (2.88)</td>
<td>MD (2.25)</td>
<td>MA (2.13)</td>
<td>HI (1.43)</td>
<td>NN (1.33)</td>
</tr>
<tr>
<td>7</td>
<td>Wenn wir die Maschine anschalten,...</td>
<td>MF (2.80)</td>
<td>HI (2.10)</td>
<td>MD (1.93)</td>
<td>MA (1.60)</td>
<td>NN (1.58)</td>
</tr>
<tr>
<td>8</td>
<td>Es regnete soviel, daß der Fuß...</td>
<td>MD (3.10)</td>
<td>MA (3.05)</td>
<td>NN (1.80)</td>
<td>MF (1.45)</td>
<td>HI (0.60)</td>
</tr>
</tbody>
</table>

Table 3.11: Total result, scale 0–4.

<table>
<thead>
<tr>
<th>approach</th>
<th>mean value</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>copy contour/natural durations</td>
<td>2.48</td>
<td>0.45</td>
</tr>
<tr>
<td>MFGI</td>
<td>2.35</td>
<td>0.52</td>
</tr>
<tr>
<td>copy contour/synthetic durations</td>
<td>2.08</td>
<td>0.49</td>
</tr>
<tr>
<td>NN modified</td>
<td>1.58</td>
<td>0.50</td>
</tr>
<tr>
<td>Hirschfeld</td>
<td>1.48</td>
<td>0.54</td>
</tr>
</tbody>
</table>

The ranking of approaches varies depending on the particular sentence, a result which might be due to the varying segmental complexity.

On the average (see Table 3.11), the samples with copy contour and natural segment durations were rated best and clearly preferred to those with copy contours and synthetic durations. MFGI is rated nearly equally well as the former. Different from the first preliminary experiment, the (modified) NN approach now outperforms the rule-based Hirschfeld approach, and in the case of sentence 3, even leads the ranking. The naturalness judgement for this approach no longer varies more strongly than that for the rule-based approaches which can be seen from the standard deviation in comparison with Table 3.9.

Figure 3.7 gives an overview of the total judgement per approach. The pie charts display the decisions in a total number of 640 (20 subjects × 4 pairs per sentence × 8 sentences) pair-comparisons per approach (equals 100 %). It can be seen, for instance, that the NN was rated superior in 28.1 % of pair-comparisons, in 22.8 % as equally good, and in 49.1 % as inferior. We observe that, quite similarly for all approaches, ‘equally good’-judgements amount to about one quarter of decisions. The percentage of these decisions reaches 25.1 % in unskilled subjects, which is only slightly higher than in skilled subjects (23.5 %).

The fact that the wide $F_0$ range in some samples affected the segmental quality of the PSOLA
Figure 3.7: Overview of all pair-comparisons per approach. Coding: approach assessed as superior (white), as equally good (gray) and inferior (black).

synthesis which only permits limited $F_0$ variations (ca. ± 30 %) might explain that the samples with copy contour and natural durations were not unanimously rated best, and that samples with only copy contour fell behind the fully synthetic ones.

The results suggest that the limited amount of material examined does not suffice for a final judgement. It becomes clear that the close approximation of natural segment durations yields a considerable improvement. In the context of a speech synthesis system this means that a poor duration control limits the naturalness achieved with a 'good' $F_0$ approach, which obviously cannot compensate for this negative effect. This trend would most probably be confirmed on a larger corpus.

3.4 Comparison of Complete TTS Systems

The objective of the main experiment was to determine how DRESS with MFGI compared with other current high-quality TTS systems for German. Besides, these TTS systems were to be compared with recorded speech.

Samples of four high-quality TTS systems for German were retrieved from interactive websites. Preconditions for choosing systems to take part in the evaluation were, that the sound inventory had been produced by male speakers and that the segmental quality was comparable to DRESS. One system was excluded due to its extremely monotonous intonation. The natural speech samples were uttered by the speaker who had produced the inventory for DRESS.

Eventually three external systems — henceforth called 'alien systems A, B, and C' — were selected for the experiment:

Two different designs were applied in this experiment:

1. a pair-wise comparison of isolated sentences as in the preceding experiments

2. an exploratory system comparison.
3.4. COMPARISON OF COMPLETE TTS SYSTEMS

Table 3.12: Approaches compared in the main experiment.

<table>
<thead>
<tr>
<th>approach</th>
<th>synthesis approach</th>
<th>intonation</th>
<th>segment durations</th>
</tr>
</thead>
<tbody>
<tr>
<td>recorded natural speech</td>
<td>-</td>
<td>natural</td>
<td>natural</td>
</tr>
<tr>
<td>DRESS</td>
<td>TD-PSOLA</td>
<td>synthetic MFGI</td>
<td>natural</td>
</tr>
<tr>
<td>DRESS</td>
<td>TD-PSOLA</td>
<td>synthetic MFGI</td>
<td>synthetic</td>
</tr>
<tr>
<td>DRESS</td>
<td>TD-PSOLA</td>
<td>synthetic NN</td>
<td>synthetic</td>
</tr>
<tr>
<td>alien system A</td>
<td>TD-PSOLA</td>
<td>synthetic</td>
<td>synthetic</td>
</tr>
<tr>
<td>alien system B</td>
<td>LPC</td>
<td>synthetic</td>
<td>synthetic</td>
</tr>
<tr>
<td>alien system C</td>
<td>TD-PSOLA</td>
<td>synthetic</td>
<td>synthetic</td>
</tr>
</tbody>
</table>

3.4.1 Pair-Comparison of Isolated Sentences

Analogous to the experiments discussed in Sections 3.3.7 and 3.3.8 a pair-wise comparison of synthesis examples of isolated sentences was conducted initially. The evaluation was performed accordingly. In every pair-comparison one point was assigned to the superior sample, and none to the inferior one. If both were assessed as equally natural, each of them received half a point. Since each system was compared with five others, the maximum score is 5.

Figure 3.8 to Figure 3.19 display two samples for each of the approaches.

The experiment was conducted with a subset of the corpus containing 15 sentences, 6 of which were declaratives with one clause, 6 declaratives consisting of two clauses, and 3 yes/no questions:

1. Statement intonation
   (single-clause sentence)
   1. In dem Lehrbuch sind viele Hinweise enthalten.
   2. Üben und immer wieder üben ist beim Erlernen jeder fremden Sprache notwendig.
   3. Das Gespräch zeigte die Gegensätze und die gemeinsamen Züge unserer Auffassungen.
   5. Wir lachten, sangen und tanzten vor Freude.
Table 3.13: System-wise comparison, rows correspond to system A, columns to system B. Cells denote vote ‘system A preferred’ values in percent.

<table>
<thead>
<tr>
<th></th>
<th>MFGI/DRESS</th>
<th>natural speech</th>
<th>alien system B</th>
<th>alien system C</th>
<th>DRESS/NN</th>
<th>alien system A</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFGI/DRESS</td>
<td>-</td>
<td>11</td>
<td>42</td>
<td>57</td>
<td>60</td>
<td>54</td>
</tr>
<tr>
<td>natural speech</td>
<td>89</td>
<td>-</td>
<td>87</td>
<td>92</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td>alien system B</td>
<td>58</td>
<td>13</td>
<td>-</td>
<td>59</td>
<td>67</td>
<td>62</td>
</tr>
<tr>
<td>alien system C</td>
<td>43</td>
<td>8</td>
<td>41</td>
<td>-</td>
<td>58</td>
<td>51</td>
</tr>
<tr>
<td>DRESS/NN</td>
<td>40</td>
<td>8</td>
<td>33</td>
<td>42</td>
<td>-</td>
<td>41</td>
</tr>
<tr>
<td>alien system A</td>
<td>46</td>
<td>8</td>
<td>38</td>
<td>49</td>
<td>59</td>
<td>-</td>
</tr>
</tbody>
</table>

II. Non-terminal Intonation
   (two clause)
7. Du muß alle kennen, wenn Du sie beurteilen willst.
8. Wir fragen ihn gern um Rat, denn er weiß meistens eine Lösung.
9. Ich weiß wirklich nicht, ob das die richtige Methode ist.
10. Wir stimmen dem Vorschlag zu, weil Ihr uns überzeugt habt.
11. Er arbeitete unermüdlich, um die Höhe seiner Kunst zu erreichen.
12. Es regnete soviel, daß der Fluß über die Ufer trat.

III. Question intonation
13. Verstehst Du mich?
14. Ist es kalt draußen?
15. Haben Sie es gehört?

As expected, the natural speech samples were rated best, but did not reach the theoretical maximum score of 5.0 (see Table 3.14). This might be due to the fact, that subjects had been requested explicitly to judge the intonational quality and in some cases arrived at an ‘equal’-judgement. Alien system B was rated slightly better than the other TTS systems, DRESS/NN lowest.

The sentence-wise results in general show a great variation in the TTS systems’ ranking depending on the particular sentence (Table 3.15). If one, however, considers the ratings for the individual sentences within one system and compares these, it becomes clear, that most sentences (1, 2, 3, 7, 9, ...) lie close to the overall mean score of the particular system, and only few sentences (6 and 10, for instance) were judged extremely good or bad respectively.

When assessing the performance of alien system C, it must be taken into account, that it does not realize any (rising) question intonation. Table 3.15 shows the results per sentence, and it can be seen that the ratings for the yes/no-questions (sentences 13-15) for system C are considerably lower than for the declarative examples. If one excludes the yes/no-questions, system C ($\mu/\sigma$: 2.27/0.50) rates slightly better than DRESS/MFGI ($\mu/\sigma$: 2.21/0.67), Table 3.16.

The good ratings for LPC-based alien system B may be partly explained by the fact, that it uses an $F_0$-range ($F_{0\text{max}}/F_{0\text{min}}$) quite close to that of natural speech and hence sounds more ‘lively’ (see Table 3.17). All PSOLA-based systems exhibit smaller $F_0$ ranges, with alien system A using very flat $F_0$ contours (compare Figures 3.10, 3.11).

The main outcome of the sentence-wise comparison is that:

- the naturalness scores for the synthesis systems DRESS with MFGI, alien systems A, B
Figure 3.8: Sentence 1 “In dem Lehrbuch sind viele Hinweise enthalten” — “The textbook contains many pieces of advice.” produced with DRESS and MFGI.

Figure 3.9: Sentence 3 “Das Gespräch zeigte die Gegensätze und die gemeinsamen Züge unserer Auffassungen.” — “The conversation showed the differences and common grounds of our opinions.” produced with DRESS and MFGI.
In dem Lehrbuch sind viele Hinweise enthalten.

Das Gespräch zeigte die Gegensätze und die gemeinsamen Züge unserer Auffassungen.

Figure 3.10: Sentence 1 produced with alien system A.

Figure 3.11: Sentence 3 produced with alien system A.
3.4. COMPARISON OF COMPLETE TTS SYSTEMS

Figure 3.12: Sentence 1 produced with alien system B.

Figure 3.13: Sentence 3 produced with alien system B.
Indem Lehrbuch sind viele Hinweise enthalten.

Figure 3.14: Sentence 1 produced with alien system C.

Das Gespräch zeigte die Gegensätze und gemeinsamen Züge unserer Auffassungen.

Figure 3.15: Sentence 3, produced with alien system C.
Figure 3.16: Sentence 1 produced with DRESS/NN.

Figure 3.17: Sentence 3 produced with DRESS/NN.
In dem Lehrbuch sind viele Hinweise enthalten.

Das Gespräch zeigte die Gegensätze und gemeinsamen Züge unserer Auffassungen.

Figure 3.18: Sentence 1, natural speech.

Figure 3.19: Sentence 3, natural speech.
3.4. COMPARISON OF COMPLETE TTS SYSTEMS

Table 3.14: Total judgement over all sentences (1-15), scale 0-5. The right-most column displays the perceptually scaled values with DRESS/MFGI as a reference for better comparison with Table 3.9.

<table>
<thead>
<tr>
<th>approach</th>
<th>mean value</th>
<th>standard deviation</th>
<th>perceptual scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>natural speech</td>
<td>4.54</td>
<td>.14</td>
<td>-1.31</td>
</tr>
<tr>
<td>alien system B</td>
<td>2.52</td>
<td>.58</td>
<td>- .21</td>
</tr>
<tr>
<td>DRESS/MFGI</td>
<td>2.27</td>
<td>.60</td>
<td>.00</td>
</tr>
<tr>
<td>alien system C</td>
<td>2.00</td>
<td>.43</td>
<td>.16</td>
</tr>
<tr>
<td>alien system A</td>
<td>1.99</td>
<td>.31</td>
<td>.17</td>
</tr>
<tr>
<td>DRESS/NN</td>
<td>1.67</td>
<td>.45</td>
<td>.38</td>
</tr>
</tbody>
</table>

Table 3.15: Sentence-wise judgement.

<table>
<thead>
<tr>
<th>No.</th>
<th>Text</th>
<th>DRESS/ MFGI</th>
<th>natural speech</th>
<th>alien system B</th>
<th>alien system C</th>
<th>DRESS/ NN</th>
<th>alien system A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>In dem Lehrbuch...</td>
<td>1.87</td>
<td>3.47</td>
<td>1.15</td>
<td>1.85</td>
<td>1.58</td>
<td>1.18</td>
</tr>
<tr>
<td>2</td>
<td>Uben und immer...</td>
<td>1.82</td>
<td>3.42</td>
<td>2.15</td>
<td>1.67</td>
<td>1.97</td>
<td>1.15</td>
</tr>
<tr>
<td>3</td>
<td>Das Gespräch...</td>
<td>1.78</td>
<td>3.08</td>
<td>2.02</td>
<td>1.25</td>
<td>1.58</td>
<td>1.42</td>
</tr>
<tr>
<td>4</td>
<td>Zangen. Pfeiler...</td>
<td>1.65</td>
<td>3.37</td>
<td>1.43</td>
<td>1.62</td>
<td>1.03</td>
<td>2.03</td>
</tr>
<tr>
<td>5</td>
<td>Wir lachten...</td>
<td>1.88</td>
<td>3.08</td>
<td>1.72</td>
<td>1.90</td>
<td>1.70</td>
<td>.95</td>
</tr>
<tr>
<td>6</td>
<td>Wir betrachten...</td>
<td>1.38</td>
<td>3.10</td>
<td>1.80</td>
<td>2.13</td>
<td>1.70</td>
<td>1.05</td>
</tr>
<tr>
<td>7</td>
<td>Du mußt alle...</td>
<td>1.67</td>
<td>3.38</td>
<td>1.93</td>
<td>1.53</td>
<td>1.08</td>
<td>1.73</td>
</tr>
<tr>
<td>8</td>
<td>Wir fragen ihm...</td>
<td>1.70</td>
<td>3.22</td>
<td>1.38</td>
<td>1.90</td>
<td>.97</td>
<td>1.77</td>
</tr>
<tr>
<td>9</td>
<td>Ich weiß wirklich...</td>
<td>1.75</td>
<td>3.40</td>
<td>2.13</td>
<td>1.53</td>
<td>1.15</td>
<td>1.23</td>
</tr>
<tr>
<td>10</td>
<td>Wir stimmen dem...</td>
<td>.92</td>
<td>3.50</td>
<td>2.07</td>
<td>2.07</td>
<td>.93</td>
<td>1.42</td>
</tr>
<tr>
<td>11</td>
<td>Er arbeitete...</td>
<td>1.30</td>
<td>3.25</td>
<td>1.63</td>
<td>1.53</td>
<td>1.47</td>
<td>1.68</td>
</tr>
<tr>
<td>12</td>
<td>Es regnete soviel...</td>
<td>1.45</td>
<td>3.35</td>
<td>1.42</td>
<td>1.32</td>
<td>1.92</td>
<td>1.45</td>
</tr>
<tr>
<td>13</td>
<td>Verstehst Du mich?</td>
<td>2.10</td>
<td>3.60</td>
<td>2.55</td>
<td>.82</td>
<td>.55</td>
<td>1.58</td>
</tr>
<tr>
<td>14</td>
<td>Ist es kalt draußen?</td>
<td>1.92</td>
<td>3.38</td>
<td>2.50</td>
<td>.38</td>
<td>.98</td>
<td>1.80</td>
</tr>
<tr>
<td>15</td>
<td>Haben Sie es gehört?</td>
<td>1.63</td>
<td>3.65</td>
<td>2.27</td>
<td>.92</td>
<td>.77</td>
<td>1.73</td>
</tr>
</tbody>
</table>

Table 3.16: Total judgement without interrogatives (sentences 1-12), scale 0-5.

<table>
<thead>
<tr>
<th>approach</th>
<th>mean value</th>
<th>standard deviation</th>
<th>perceptual scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>natural Speech</td>
<td>4.49</td>
<td>.46</td>
<td>-1.72</td>
</tr>
<tr>
<td>alien system B</td>
<td>2.33</td>
<td>.71</td>
<td>- .14</td>
</tr>
<tr>
<td>alien system C</td>
<td>2.27</td>
<td>.50</td>
<td>- .08</td>
</tr>
<tr>
<td>DRESS/MFGI</td>
<td>2.21</td>
<td>.67</td>
<td>.00</td>
</tr>
<tr>
<td>alien system A</td>
<td>1.91</td>
<td>.34</td>
<td>.18</td>
</tr>
<tr>
<td>DRESS/NN</td>
<td>1.81</td>
<td>.52</td>
<td>.23</td>
</tr>
</tbody>
</table>
and C are quite closely clustered

- the quality distance between these systems and natural speech is still great and corresponds to almost half of the scale used.

### 3.4.2 System Comparison

The objective of the system comparison was to assess subjects’ acceptance for each of the TTS systems taking part in the evaluation. All preceding experiments based on pair-wise comparison had shown that the perceived naturalness strongly depended on the particular sentences assessed. Besides, full speech synthesis in practice is employed for reading coherent texts, whereas announcements in public transport, for instance, are generally produced by concatenating larger chunks of recorded natural speech. For these reasons, the following coherent news text was chosen and synthesized with the approaches to be compared:

“Hell erleuchtet präsentierte sich am späten Montag abend die Kuppel des Reichstags. Die Bundestags-Baukommission sah sich vor Ort an, wie das künftige Parlamentsgebäude einmal wirken soll. Zu besonderen Anlässen wie der Wahl des Bundespräsidenten soll die Kuppel effektiv leuchten.” — “Late Monday evening the cupola of the Reichstag presented itself brightly illuminated. The construction commission of the Bundestag examined in place, what the future parliament building will look like. On special occasions, such as the election of the Federal President, the cupola will be illuminated effectively.” (taken from: Berliner Zeitung, June 16, 1998)

The experiment was conducted using the program RANKING described in Section 3.2.2. In addition to the approaches compared in the preceding experiment, a version of MFGI/DRESS was employed, which featured natural segment durations extracted from read renderings of the news text. Figures 3.20 to 3.24 display the first sentence of the text created with different TTS systems.

Many subjects commented that the request to assign 0 points to the least acceptable TTS system posed problems, not necessarily because the distance between systems could not be expressed by the number of points, but because of the negative connotation of ‘0 points’, since none of the systems was perceived as being completely unacceptable. Less skilled subjects observed difficulties in abstracting from the segmental impression and concentrating on the prosodic quality. Hence we cannot exclude that in the case of prosodically equally acceptable examples judgements were influenced by the quality or the pleasantness of the synthetic voice.

As could have been expected, the natural speech samples were rated best and unanimously reached the maximum score of 10.0 (see Figure 3.25 for mean scores). The best TTS systems, including DRESS with MFGI cluster around a mean value of 5.0, whereas alien system A falls
Figure 3.20: Sentence 1 from the news text produced with DRESS/MFGI.

Figure 3.21: Sentence 1 from the news text produced with alien system A.
Figure 3.22: Sentence 1 from the news text produced with alien system C.

Figure 3.23: Sentence 1 from the news text produced with DRESS/NN.
Figure 3.24: Sentence 1 from the news text, natural speech.

Figure 3.25: Total judgement, system comparison.
Table 3.18: Total result (scale: 0–10 points).

<table>
<thead>
<tr>
<th>approach</th>
<th>mean value</th>
<th>standard deviation</th>
<th>frequency &quot;0 points&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>natural speech</td>
<td>10.00</td>
<td>.00</td>
<td>0</td>
</tr>
<tr>
<td>DRESS/MFGI, natural durations</td>
<td>6.57</td>
<td>2.02</td>
<td>1</td>
</tr>
<tr>
<td>DRESS/MFGI</td>
<td>5.35</td>
<td>2.60</td>
<td>2</td>
</tr>
<tr>
<td>alien system C</td>
<td>4.83</td>
<td>2.55</td>
<td>3</td>
</tr>
<tr>
<td>DRESS/NN</td>
<td>4.70</td>
<td>2.58</td>
<td>3</td>
</tr>
<tr>
<td>alien system B</td>
<td>4.43</td>
<td>2.69</td>
<td>4</td>
</tr>
<tr>
<td>alien system A</td>
<td>2.65</td>
<td>2.74</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 3.19: Results from paired T-test for significance, h.s. = highly significant, s. = significant, n.s. = not significant.

<table>
<thead>
<tr>
<th></th>
<th>DRESS/ MFGI n.d.</th>
<th>DRESS/ MFGI</th>
<th>alien system C</th>
<th>DRESS/ NN</th>
<th>alien system B</th>
<th>alien system A</th>
</tr>
</thead>
<tbody>
<tr>
<td>natural speech</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>DRESS/ MFGI n.d.</td>
<td>.011</td>
<td>.023</td>
<td>.005</td>
<td>.018</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>alien system C</td>
<td>.514</td>
<td>.345</td>
<td>.272</td>
<td>.009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alien system B</td>
<td></td>
<td>.878</td>
<td>.572</td>
<td></td>
<td>.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.752</td>
<td></td>
<td>.023</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.039</td>
<td></td>
</tr>
</tbody>
</table>

off considerably. The DRESS/MFGI variant with natural segment durations was preferred to the all-synthesis approaches and reached 6.6 points.

The frequency with which a system was rated least acceptable corresponds to the mean preference score (Table 3.18).

Paired-samples T-test was performed and showed that the results stated so far are significant ($p < .05$) or highly significant ($p < .01$). Table 3.19 gives detailed results.

The total number of points assigned in the experiment varied depending on the subject ($\mu = 38.5, \sigma = 5.1$). One subject gave extremely low ratings to all synthetic examples ($\mu = 23$).

3.4.3 Discussion

First of all it must be stated that the text sample used in this experiment is very short and was selected ad-hoc, and that different material might have produced different results.

The ranking of TTS systems found in the present experiment differs from the one observed in the sentence-wise pair-comparison, where alien system B had significantly outperformed the
other TTS systems. We had then concluded that this result might be explained by the fact that the $F_0$ range used by system B was wider than that of the other systems and equaled the range found in natural speech, making system B sound more lively. In the current experiment on a news passage a wide $F_0$ range might have had the opposite effect, being perceived as overly emotional.

We believe that the design applied in the present study is more realistic than a pair-wise comparison of isolated utterances, since TTS systems are typically used for reading coherent passages of text, i.e. news bulletins, traffic information, weather reports, made up of more than a single sentence. If a potential customer wished to select a system for his purposes, he would best be presented with an interactive dialog similar to the one used in this study permitting him to play back the same text passage with each of the TTS systems available. We conclude that:

- the synthesis systems DRESS with MFGI and alien systems B and C are perceived as being almost equally natural,
- the quality distance between these systems and natural speech is still great and corresponds to almost half of the scale used,
- the fact that the DRESS/MFGI variant with natural durations was found more acceptable than the full synthesis systems indicates that modeling natural segment durations can considerably improve the perceived prosodic quality

### 3.5 Discussion and Conclusions

The series of perception experiments documented in this chapter leads to the following conclusions:

- MFGI is preferred to other approaches to $F_0$ modeling tested in the DRESS environment.
- DRESS equipped with MFGI performs as well as other presently available high-quality TTS systems for German.
- TTS systems are still rated as being of poor prosodic quality compared with recorded speech.
- The poor quality of the current duration model in DRESS adversely affects the prosodic quality of the synthetic speech and cannot be compensated by the quality of $F_0$ prediction.

Since the $F_0$ contour is aligned with respect to the segment durations computed by the duration model, gross duration errors directly affect the perceived prosodic naturalness. This is indicated by the better quality of stimuli employing durations copied from natural utterances. Hence, the key to improved prosodic quality lies in the better modeling of the interaction between $F_0$ and duration. To this effect, the following chapter documents three preliminary studies.
Chapter 4

$F_0$ and Duration - Preliminary Studies

Abstract
In this chapter we document three preliminary studies of the relationship between the $F_0$ contour and the segments and their durations. The studies are focussed on the following topics: (1) The timing of accent commands with respect to the syllabic structure, (2) the perception of prominence depending on syllabic duration and accent command amplitude, and (3) the relationship between focus and sentence mode and the $F_0$ and duration contours.
4.1 Introduction

The studies documented in this chapter were designed as preliminary steps towards an improved prosodic model predicting $F_0$ as well as duration contours.

Experiments 1 and 3 are based on specially designed small corpora which permitted a controlled and detailed analysis of the interaction between $F_0$ and duration. The first study (Section 4.2) concerns the relationship between the syllabic structure and the timing of accent commands underlying the $F_0$ contour of the utterance. This temporal relationship is especially important for deciding how to link accent commands to syllables.

Section 4.3 is dedicated to the contributions of $F_0$ and syllable duration to perceived syllable prominence. As the relative prominence of constituents of an utterance is directly related to the message conveyed, an important question is how the production-based Fujisaki model parameter $A_R$ is correlated with perceived prominence.

Section 4.4 examines how the factors focus, sentence mode and the presence of boundaries influence the $F_0$ and duration contours of an utterance.

4.2 The Influence of the Syllable Structure on the $F_0$ contour

4.2.1 Introduction

In a short study we examined the influence of syllable structure on the onset and offset times of accent commands of the Fujisaki model. On the example of three-syllable words with word-accent on the second which were uttered in isolation, it was examined which factors influence accent command onset time $T_1$ and accent command offset time $T_2$. $T_1$ can be predicted accurately either with respect to the syllable onset or the onset of the nuclear vowel, provided, the internal timing of the syllable is known. Otherwise, the syllable onset is the more appropriate reference. $T_1$ is also influenced by the type of the consonant in the onset of the syllable. $T_2$ basically coincides with the segmental offset of the syllable.

As explained in Section 2.7.3, the temporal alignment of accent commands in MFGI is performed with respect to the onset of the nuclear vowel of the accented syllable. This yields good results in certain syllable environments such as voiced consonant plus long vowel, but in cases of short vowels preceded by unvoiced consonant clusters, for instance, the accent is perceived as too weak and in some cases even sounds unnatural as it appears to be shifted towards the following syllable.

The study presented in this section examined more closely the fine temporal alignment of accent commands and accented syllables.

4.2.2 Speech Material

The speech material used in the study consists of utterances of three-syllable words (mostly verbs) by three native speakers of German, twice each. The words, altogether 67 tokens, exhibit the structure 'be'-accented syllable-plosive-en', for instance: ‘be-den-ten’ — ‘to signify’, ‘be-fragen’ — ‘to interrogate’, ‘be-stri-tten’ — ‘denied’ etc. This kind of material was chosen in order to facilitate syllabic segmentation.

Table 4.1 gives an overview of syllable structures examined. Considering the great diversity of syllable structures in German this selection is far from complete.

In order to assess the microprosodic influence of the sounds involved on the $F_0$ contour, all tokens were uttered monotonously once.
4.2.3 Method of Analysis

The utterances were recorded on tape and converted at 16 kHz/16 bit. The $F_0$ contours were extracted and analyzed using the Fujisaki model by the method of Analysis-by-Synthesis. The speech segments of the accent syllables were labeled auditorily.

Figure 4.1 shows an example of analysis of the word ‘bemänteln’ — ‘to cover up’. At the top the speech waveform is displayed. The curve drawn using + symbols indicates the measured $F_0$ contour, the solid line the $F_0$ contour produced by the Fujisaki model and the dashed line its phrase component. The underlying accent command is displayed at the bottom. The vertical lines mark the boundaries of sound segments belonging to the word-accent syllable which have been SAMPA labeled.

The $F_0$ contours of all tokens can be modeled using a single accent command and a single phrase command.

Figure 4.2 shows examples of accented and monotonous version of the word ‘bewegten’ — ‘moved’ by speaker HM. As can be seen from the monotonous version, the $F_0$ contour exhibits a downward trend at the [v]. In contrast, the $F_0$ contour on ‘en’ does not show any influence by the preceding sound [t] in the monotonous version, hence the falling slope in the accented version is most likely caused by the falling slope of the accent command. This kind of effects must be taken into account when modeling the $F_0$ contours of the accented tokens, since the Fujisaki model does not consider micropragmatic influences.

4.2.4 Results of Analysis

Possible options for the alignment of accent command onsets and offsets with accented syllables include the alignment (1) relative to accent syllable durations, (2) relative to nuclear vowel durations and (3) absolute timing, i.e. the alignment with vowel or syllable onsets or offsets.

The duration of the consonant cluster preceding the nuclear vowel and the type of consonant immediately preceding the nuclear vowel were examined as further factors.
Figure 4.2: Examples of microprosodic effects on the $F_0$ contours: Word ‘bewegen’ — ‘moved’, dips caused by the sound [v]. Top: Monotonous version, bottom: Regular citation.
Figure 4.3: Examples of synthetic $F_0$ contours. Top: ‘bestatten’ — ‘to bury’, bottom: ‘bemühren’ — ‘constrained’.
4.2. THE INFLUENCE OF THE SYLLABLE STRUCTURE ON THE F₀ CONTOUR

Table 4.1: Syllable structures examined in the experiment.

<table>
<thead>
<tr>
<th>syllable structure</th>
<th>number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>1</td>
</tr>
<tr>
<td>C/V</td>
<td>8</td>
</tr>
<tr>
<td>C/C/V</td>
<td>10</td>
</tr>
<tr>
<td>C/C/C/V</td>
<td>1</td>
</tr>
<tr>
<td>V/C</td>
<td>2</td>
</tr>
<tr>
<td>C/V/C</td>
<td>31</td>
</tr>
<tr>
<td>C/C/V/C</td>
<td>9</td>
</tr>
<tr>
<td>V/C/C</td>
<td>1</td>
</tr>
<tr>
<td>C/V/C/C</td>
<td>3</td>
</tr>
<tr>
<td>C/C/V/C/C</td>
<td>1</td>
</tr>
</tbody>
</table>

4.2.4.1 Accent Command Onset Time T1

In order to test the possible alignment options, the following parameters were determined:

- \( t_{son} \): segmental onset time of the accented syllable
- \( t_{vcon} \): onset time of the nuclear vowel
- \( dur_s \): syllable duration
- \( dur_v \): nuclear vowel duration
- \( dur_{on} \): duration of syllable onset
- \( T1_{dist} = T1 - t_{son} \): the delay between the onset time of the nuclear vowel and T1
- \( T1_{distv} = T1 - t_{vcon} \): the delay between the onset time of the nuclear vowel and T1
- \( T1_{r,s} \): \( T1_{dist} \) expressed as a fraction of the accent syllable duration
- \( T1_{r,v} \): \( T1_{distv} \) expressed as a fraction of the nuclear vowel duration

Mean values and standard deviation for some of these parameters are displayed in Table 4.2.

Table 4.2: Mean values of some of the parameters extracted.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T1_{dist} )</td>
<td>218 ms</td>
<td>71 ms</td>
</tr>
<tr>
<td>( T1_{distv} )</td>
<td>-58 ms</td>
<td>58 ms</td>
</tr>
<tr>
<td>( T1_{r,s} )</td>
<td>51 %</td>
<td>11 %</td>
</tr>
<tr>
<td>( T1_{r,v} )</td>
<td>26 %</td>
<td>25 %</td>
</tr>
<tr>
<td>( dur_s )</td>
<td>420 ms</td>
<td>80 ms</td>
</tr>
<tr>
<td>( dur_v )</td>
<td>180 ms</td>
<td>80 ms</td>
</tr>
<tr>
<td>( dur_{on} )</td>
<td>160 ms</td>
<td>60 ms</td>
</tr>
</tbody>
</table>

Correlations between \( T1_{dist} \), \( T1_{distv} \) and the durations of syllable, syllable onset and nuclear vowel were calculated which are listed in Table 4.3.
Table 4.3: Correlations between $T_{1\text{dist}}$ and $T_{1\text{distv}}$ and various parameters. ‘$n_{on}$’ denotes the number of consonants in the syllable onset.

<table>
<thead>
<tr>
<th>parameter $p$</th>
<th>$\rho(T_{1\text{dist}}, p)$</th>
<th>$\rho(T_{1\text{distv}}, p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$dur_s$</td>
<td>.73</td>
<td>-.42</td>
</tr>
<tr>
<td>$dur_v$</td>
<td>.42</td>
<td>-.76</td>
</tr>
<tr>
<td>$dur_{on}$</td>
<td>.62</td>
<td>.28 ($p = .01$)</td>
</tr>
<tr>
<td>$n_{on}$</td>
<td>.44</td>
<td>.31</td>
</tr>
</tbody>
</table>

Regression models for $T_{1\text{dist}}$ and $T_{1\text{distv}}$ were built and tested on the data in the corpus. The correlation between measured and predicted $T_{1\text{dist}}$ and $T_{1\text{distv}}$ was calculated. Table 4.4 shows the results. It can be seen that the prediction with respect to the syllable onset yields better correlations, but the RMSE for both alignment options is the same. This indicates that both alignment options are practically equivalent, provided the precise internal timing of the syllable is known.

Table 4.4: Correlations and root mean square error for predicting $T_{1}$.

<table>
<thead>
<tr>
<th>alignment</th>
<th>$\rho(\text{pred., obs.})$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>abs. from syllable onset</td>
<td>.88</td>
<td>31 ms</td>
</tr>
<tr>
<td>abs. from vowel onsets</td>
<td>.82</td>
<td>31 ms</td>
</tr>
</tbody>
</table>

If only the total duration of the syllable and the number of consonants in the onset is known, $T_{1\text{dist}}$ proves to be a better alignment parameter for the accent command onset than $T_{1\text{distv}}$. A regression model with $dur_s$ and $n_{on}$ as the input parameters yields the results shown in Table 4.5.

Table 4.5: Correlations and root mean square error for predicting $T_{1}$ with a simpler regression model.

<table>
<thead>
<tr>
<th>alignment</th>
<th>$\rho(\text{pred., obs.})$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>abs. from syllable onset</td>
<td>.78</td>
<td>45 ms</td>
</tr>
<tr>
<td>abs. from vowel onsets</td>
<td>.59</td>
<td>47 ms</td>
</tr>
</tbody>
</table>

Predicting $T_{1}$ relative to syllable or vowel duration (parameters $T_{1\text{rad}}$ and $T_{1\text{rad}}$) yielded considerably poorer results.

Confining the analysis on single-consonant onsets, we examined the dependency of $T_{1\text{dist}}$ on the type of the preceding consonant. In order to account for the different mean durations of phones we calculated the ratio between $T_{1\text{dist}}$ and the onset, i.e. the consonant duration. It is found that, though most consonant type-specific mean are close to the total mean value, some differ considerably. These are listed in Table 4.6.
4.2. THE INFLUENCE OF THE SYLLABLE STRUCTURE ON THE F0 CONTOUR

Table 4.6: Ratio between \( T_{1,\text{dist}} \) and consonant duration. Small values correspond to relatively early onsets of accent commands.

<table>
<thead>
<tr>
<th>Consonant</th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>[S]</td>
<td>.77</td>
<td>.13</td>
</tr>
<tr>
<td>[k]</td>
<td>1.13</td>
<td>.10</td>
</tr>
<tr>
<td>[R]</td>
<td>1.19</td>
<td>.52</td>
</tr>
<tr>
<td>mean</td>
<td>1.39</td>
<td>.36</td>
</tr>
<tr>
<td>[l]</td>
<td>1.65</td>
<td>.31</td>
</tr>
<tr>
<td>[j]</td>
<td>1.69</td>
<td>.24</td>
</tr>
<tr>
<td>[m]</td>
<td>1.75</td>
<td>.29</td>
</tr>
</tbody>
</table>

4.2.4.2 Accent Command Offset Time \( T2 \)

The following parameters were calculated in order to determine the factors influencing \( T2 \):

\[
T_{2,\text{dist}} = T2 - t_{o\text{sf}} \quad \text{the delay between the offset time of the syllable and measured } T2 \\
T_{2,\text{dist,v}} = T2 - t_{v.o\text{sf}} \quad \text{the delay between the offset time of the nuclear vowel and measured } T2 \\
T_{2_{r,\text{sd}}} \quad \text{measured } T2 \text{ expressed as a fraction of the accent syllable duration} \\
T_{2_{r,\text{rel}}} \quad \text{measured } T2 \text{ expressed as a fraction of the nuclear vowel duration}
\]

In line with the considerations concerning \( T1 \), similar regression models for predicting \( T2 \) were derived and tested. It must be stated, however, that in some cases, where the accent syllable-final consonants are voiceless, \( T2 \) cannot be determined exactly, unless inferred from the portion of the \( F0 \) contour on the final syllable, as, for instance, in Figure 4.3. Table 4.7 shows the correlation coefficients for \( T_{2,\text{dist}} \) and \( T_{2,\text{dist,v}} \) with respect to the durational properties of the syllable. It becomes clear that the onset of the syllable has practically no influence on the timing of \( T2 \).

Table 4.7: Correlations between \( T_{2,\text{dist}} \) and \( T_{2,\text{dist,v}} \) and various parameters.

<table>
<thead>
<tr>
<th>parameter ( p )</th>
<th>( \rho(T_{2,\text{dist}}, p) )</th>
<th>( \rho(T_{2,\text{dist,v}}, p) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( dur_s )</td>
<td>(-.37)</td>
<td>.09 (n.s.)</td>
</tr>
<tr>
<td>( dur_v )</td>
<td>(-.30)</td>
<td>(-.61)</td>
</tr>
<tr>
<td>( dur_{on} )</td>
<td>(-.10) (n.s.)</td>
<td>(-.09) (n.s.)</td>
</tr>
</tbody>
</table>

Regression models for \( T_{2,\text{dist}} \) and \( T_{2,\text{dist,v}} \) including the significant parameters from Table 4.7 yields considerably low correlations of 0.43 and 0.42, \( \sigma > 49 \text{ ms and 50 ms, respectively}. \) This indicates that \( T2 \) is not strongly related to the syllable’s internal structure, but rather aligned with respect to the syllable offset time proper (\( \rho(T2, t_{o\text{sf}}) = 0.891 \)).

For neither \( T1 \) nor \( T2 \) a significant correlation with the accent command amplitude \( Aa \) was found.
4.2.5 Perceptual Evaluation

For evaluating the effectiveness of the refined timing model, an informal perception experiment was conducted with synthesized stimuli produced with the TU Dresden TTS system. Figure 4.3 shows examples of model-generated $F_0$ contours for the German words ‘bestatten’ — ‘to bury’ and ‘bemühen’ — ‘constrained’ which were produced following the refined alignment rules.

Four native speakers of German were offered pairs of stimuli, one stimulus produced with the original alignment rules and one with the refined ones, and had to decide which version they found more natural or if both were equally (un)natural. A slight preference for the latter stimuli was found, especially in the aforementioned problematic cases.

Figure 4.3 shows examples of model-generated $F_0$ contours for the German words ‘bestatten’ — ‘to bury’ and ‘bemühen’ — ‘constrained’ which were produced following the refined alignment rules.

4.2.6 Discussion and Conclusions

The data used in this study is rather limited and only permits tentative conclusions. Models were derived from and tested on the same material due to the relatively small amount of samples. Segmentation inaccuracies may have further flawed the results. Besides, the speech material does not contain a complete selection of German speech sounds and sound combinations. Rules will have to be established for all types of possible leading consonants, as there is no evidence that they can be clustered into groups of similar sounds, such as nasals, unvoiced plosives, etc. The perceptual evaluation shows that considering the syllable structure when aligning the accent command offsets with accented syllables yields better results than the original ad-hoc rules. Alignment, however, may be more at variance in continuous speech than in utterances of citation forms.

Whereas $T_1$ was reliably predicted from the internal timing of the syllable, especially the duration of the syllable onset, no strong influence of the syllable structure on $T_2$ was found, as the latter basically aligns with respect to the syllable offset. Hence, the timing of accent commands is defined by the structure of the syllable, but not by the strength in terms of accent command amplitude $A_{a}$.

4.3 Perception of Prominence

Experimental studies have shown that an increase in the prominence of a speech syllable can be achieved by increasing the tone interval of the $F_0$ movement assigned to the syllable, as well as by stretching the speech segments of the syllable. On the example of synthetic stimuli it was observed that a given reduction in prominence caused by the reduction of a tone interval can be compensated by a proportional increase in syllable duration. In other words, listeners cannot clearly specify which feature modified causes a perceptual change in prominence.

The study documented in this section was carried out in cooperation with Christina Widera (University of Bonn) and investigates the relationship between perceived syllable prominence and the $F_0$ contour in terms of MFG1 [MW01]. A subcorpus of the Bonn Prosodic Database was analyzed using the $F_0$ model, and normalized log syllable durations were calculated.

4.3.1 Introduction

As indicated earlier in this thesis, one important function of prosody is the highlighting of linguistic units. Investigations show that the perceived prominence of these units can be regarded as a gradual parameter. It is suited for describing the emphasis assigned to linguistic units in
relation to their environment. This study focuses on the relationship between the perceived prominence of a syllable and two important prosodic features assigned to the syllable. These features are (1) the interval of a major $F_0$ transition connected to the syllable as yielded by parametrizing the $F_0$ contour with a quantitative model, (2) normalized log syllable durations.

### 4.3.1.1 Prosodic features examined in this study

As discussed in Section 2.7, in the framework of MFGI a given $F_0$ contour is described as a sequence of linguistically motivated tone switches, major rises and falls, which are modeled by onsets and offsets of accent commands connected to accented syllables, or by so-called boundary tones before prosodic boundaries. Hence the interval of a tone switch readily relates to the accent command amplitude $Aa$ assigned to it. Tone switches constitute functionally distinct intonational elements, so-called 'intonemes'. In the scope of this study, we concentrate on the classes $N^\uparrow$-intoneme ('non-terminal intoneme' at phrase-medial accents, rising tone switch), and $I^\downarrow$-intoneme ('information intoneme' at declarative-final accents, falling tone switch).

### 4.3.1.2 Prominence of syllables

The notion of prominence followed in this study is based on [FK89]. Three labelers had to judge the degree of prominence on the syllable level relative to the surrounding syllables on a scale from 0 to 31. Between subjects, the labeled prominences correlate strongly ($\rho > .80$; [HP95a]). Earlier investigations show that the relation between prominence ratings and syllable duration, as well as $F_0$ peaks, described by parameters of a maximum based description of $F_0$-contours [HP$^+$95b], are linear. However, prominence is also related to linguistic features (i.e. word class, position in a phrase, and focus). Furthermore, perceived prominence is reliably predicted from linguistic features, as well as from acoustic features [HP$^+$95b]. Thus perceived prominence can be regarded as a gradual parameter integrating linguistic features and acoustic parameters.

Since the Fujisaki model is inherently production-based, one major issue in this study is to establish the relationship between the amplitude parameter $Aa$ and the perceived prominence of a syllable. Furthermore the implicit claim underlying MFGI that not all parts of the $F_0$ contour are 'equally important' is investigated. If the claim is correct, linguistically motivated $F_0$ transitions, i.e. tone switches, should strongly contribute to the perceived prominence of a syllable, whereas so-called 'pitch-interrupters' (Isachenko), $F_0$ transitions at non-accent syllables, should not.

### 4.3.2 Speech Material and Method of Analysis

The speech material was taken from the Bonn Prosodic Database (BPD, [Heu99]). The BPD contains read speech of three German speakers. The subset is composed of isolated sentences, question-answer pairs, and short stories of one female speaker, and contains a total of 3,401 syllables. Every syllable is assigned information about its position and its number in higher-level units (i.e. position of syllable in a word or in a prosodic phrase), its nucleus, as well as the number of phones it consists of. The syllables are annotated with their word class and lexical word stress, as well as their prominence scaled from 0 to 31, as judged by three phoneticians. The prominence of a syllable is taken to be the median of the judgments. Log syllable durations were computed from phone labels in the BPD and normalized to their syllable count and the property of the nuclear vowel, being either schwa or non-schwa, the most important intrinsic features. $F_0$ contours were extracted at intervals of 10 ms, and the Fujisaki parameters determined using an automatic multi-stage approach (see Section 5.3.2 for details). All parameter sets were then checked visually as well as auditorily and errors corrected. Tone switches were assigned to
syllables by evaluating the timing of accent commands with respect to syllabic boundaries. In the case of potentially accented syllables (i.e., word accent syllables of content words) also the preceding and the following syllable were taken into account.

4.3.3 Results of Analysis

Figure 4.4 shows an example of analysis from the database displaying the utterance “Ist das die einzige Möglichkeit? — Ja, so ist es.” — “Is this the only possibility? — Yes, it is.” The figure displays from top to bottom: the speech waveform, the extracted and model-generated \(F_0\) contours, the duration contour in terms of the syllable z-score drawn as horizontal lines of the length of the respective syllable, the SAMPA transcription of the utterance, the underlying phrase and accent commands and median perceived prominence. Mean accent commands amplitude \(A_a\) and normalized log syllable duration \((n_{syl\_dur})\) depending on grouped prominence values.

4.3.3.1 Perceived Prominence and Acoustic Parameters

Perceived prominences are evaluated in relation to the acoustic parameters \(A_a\) and normalized log syllable duration \((n_{syl\_dur})\). The correlation over all syllables (Pearson correlation coefficient) is about 0.5 for \(A_a\) \((p < .01, N = 3401)\) and about 0.4 for \(n_{syl\_dur}\) \((p < .01, N = 3,399)\). These relatively low values may be explained by other influences such as phrase-final lengthening and boundary tones. If we only include syllables with lexical word accent, the correlation between
prominence values and $Aa \ (\rho = .60, p < .01, N = 2,043)$ as well as $nsyl\dur\ (\rho = .50, p < .01, N = 2,043)$ increases. However, $Aa$ is mostly related to higher prominence values ($> 15$). In contrast to $Aa$, $nsyl\dur$ is more strongly correlated with lower prominence values ($< 16$). Hence, levels of weaker prominence are mostly associated with duration and strongly perceived prominence is mostly related to $F_0$ movements. The relationship between perceived prominence and the two acoustic parameters can be regarded as nearly linear. This becomes more obvious when the prominence values are grouped (Figure 4.5).

### 4.3.3.2 Perceived Prominence and Tone Switches

In this section we examine whether the linguistic notion of tone switches is reflected by prominence values. Prominence values and acoustic parameters of the linguistically motivated tone switches ($\updownarrow$-intonemes and $\uparrow$-intonemes) are compared with the values of non-linguistic $F_0$ movements, i.e. rising and falling pitch interrupters; c.f. Figure 4.6 and Figure 4.7. The comparison of falling pitch interrupters ($N = 294$) with $\updownarrow$-intonemes ($N = 395$) yields significant differences with respect to their mean prominence values ($t = 20.80, df = 43.72, p < .01$) and their acoustic values ($Aa : t = 3.11, df = 687, p < .01; nsyl\dur : t = 4.768, df = 554.83, p < .01$). $\updownarrow$-intonemes more strongly contribute to prominence than falling pitch interrupters. Comparable results are also found for the $\uparrow$-intoneme ($N = 422$) and rising pitch interrupters ($N = 272$). $\uparrow$-intonemes are associated with stronger prominence than not linguistically motivated $F_0$ rises ($t = 11.71, df = 411.51, p < .01$).

![Graph](image)

Figure 4.5: Mean accent commands amplitude $Aa$ and normalized log syllable duration ($nsyl\dur$) depending on grouped prominence values.

However, no significant differences are established for $nsyl\dur$. Furthermore the results show that the average prominence value of $\uparrow$-intonemes ($N = 422$) is lower than those of the $\downarrow$-intonemes ($N = 395$). The significant difference between the prominence values ($t = -7.74, df = 813.40, a < .01$) is also reflected by the acoustic parameters $Aa$ ($t = -3.37, df = 788.98, a < .01$) and $nsyl\dur$ ($t = -4.96, df = 762.67, a < .01$).
Detailed analysis shows that prominences correlate with the timing of the accent command with respect to the syllable. Accented syllables of \( \downarrow \) intonemes with an accent command on- and offset in the syllable proper \((N = 135)\) are perceived more prominent \((\text{mean} = 23.5)\) than those with command on- and offset in preceding syllables \((\text{mean} = 21.5, N = 63; t = 2.89, df = 196, a < .01)\). Hence the position of the accent command with regard to the accented syllable influences prominence ratings. The result also indicates that, although the \( F_0 \) movement takes place in an unaccented syllable, the perceived prominence is assigned to the following accented syllable. In contrast, for \( N \uparrow \)-intonemes, at least in the data examined prominence judgements are not significantly affected by the timing of the accent command. In conclusion, the comparison between prominence values assigned to different classes of \( F_0 \) transitions indicates that prominence ratings reflect a linguistically motivated association and interpretation of \( F_0 \) movements.

![Figure 4.6: Mean prominence of linguistic (information intoneme (\( \downarrow \)), non-terminal intoneme (\( N \uparrow \))) and non-linguistically motivated (falling (\( \downarrow \)P) and rising pitch interrupter (\( \uparrow \)P)) \( F_0 \) transitions.](image)

### 4.3.4 Discussion and Conclusions

Analysis shows that, for accented syllables, prominences are strongly correlated with the amplitude \( A_a \) of accent commands underlying the \( F_0 \) movements in these syllables, whereas comparable \( F_0 \) movements in unaccented syllables have little effect on prominence. The influence of \( A_a \) versus syllable duration on prominence is greater in higher prominence classes. The fact that the \( F_0 \) movement does not necessarily take place in the accented syllable proper, indicates that the prominence judgment is partly guided by linguistic considerations, a finding in accordance with results of earlier studies of prominence.

We may tentatively interpret the relationship between \( A_a \) and perceived prominence as follows: While \( A_a \) - inter alia - reflects the ‘relative importance’ of accented constituent words in an utterance as intended by the speaker, prominence reflects the ‘realized performance structure’ of the utterance as perceived by the listener. In the context of TTS the results indicate that \( F_0 \)
4.4. Focus, Sentence Mode and Phrase Boundary Location

4.4.1. Introduction

The following relevant linguistic factors influencing the $F_0$ contour were identified by the author in his D.Eng. thesis [Mix98]:

1. the sentence mode
2. the word accent location
3. the type of focus (narrow vs. broad)
4. the part of speech (names and nouns receive more prominence than verbs, for instance).
5. the speech rate (increased speech rate compresses the $F_0$ range)

In order to examine the influence of factors 1 to 3 on syllable durations, a production experiment was performed which is documented in the following sections.

As discussed in Section 2.6.2 current duration models are based on the statistical evaluation of large databases. This provides for a good coverage of possible segmental environments. The influence of focal conditions, sentence mode and boundary location on syllable duration, however, may be blurred, because the segmental contexts in which they occur vary throughout the database. Besides, the relationship with the $F_0$ contour is usually neglected in this kind of analysis. The current study closely examines data where the segmental context is kept constant, but the underlying linguistic information varies.
Table 4.8: List of all contexts examined.

<table>
<thead>
<tr>
<th>No.</th>
<th>phrasal condition</th>
<th>sentence mode</th>
<th>focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>single-phrase</td>
<td>statement</td>
<td>broad</td>
</tr>
<tr>
<td>2</td>
<td>single-phrase</td>
<td>statement</td>
<td>narrow on ‘U-Bahn’</td>
</tr>
<tr>
<td>3</td>
<td>single-phrase</td>
<td>statement</td>
<td>narrow on ‘Ruhleben’</td>
</tr>
<tr>
<td>4</td>
<td>single-phrase</td>
<td>echo-question</td>
<td>broad</td>
</tr>
<tr>
<td>5</td>
<td>single-phrase</td>
<td>echo-question</td>
<td>narrow on ‘U-Bahn’</td>
</tr>
<tr>
<td>6</td>
<td>single-phrase</td>
<td>echo-question</td>
<td>narrow on ‘Ruhleben’</td>
</tr>
<tr>
<td>7</td>
<td>two-phrase-initial</td>
<td>continuation</td>
<td>broad</td>
</tr>
<tr>
<td>8</td>
<td>two-phrase-initial</td>
<td>continuation</td>
<td>narrow on ‘U-Bahn’</td>
</tr>
<tr>
<td>9</td>
<td>two-phrase-initial</td>
<td>continuation</td>
<td>narrow on ‘Ruhleben’</td>
</tr>
<tr>
<td>10</td>
<td>two-phrase-initial</td>
<td>echo-question</td>
<td>broad</td>
</tr>
<tr>
<td>11</td>
<td>two-phrase-initial</td>
<td>echo-question</td>
<td>narrow on ‘U-Bahn’</td>
</tr>
<tr>
<td>12</td>
<td>two-phrase-initial</td>
<td>echo-question</td>
<td>narrow on ‘Ruhleben’</td>
</tr>
<tr>
<td>13</td>
<td>phrase boundary after ‘U-Bahn’</td>
<td>statement</td>
<td>broad</td>
</tr>
<tr>
<td>14</td>
<td>two-phrase-final</td>
<td>statement</td>
<td>broad</td>
</tr>
<tr>
<td>15</td>
<td>two-phrase-final</td>
<td>statement</td>
<td>narrow on ‘U-Bahn’</td>
</tr>
<tr>
<td>16</td>
<td>two-phrase-final</td>
<td>statement</td>
<td>narrow on ‘Ruhleben’</td>
</tr>
</tbody>
</table>

4.4.2 Speech Material and Method of Analysis

The idea underlying the corpus design was to cover a number of different linguistic functions of prosodic cues with a small number of sentences. The target utterance should contain mostly voiced sounds ensuring a continuous $F_0$ contour. Hence, the German sentence “Wir nehmen die U-Bahn nach Ruhleben” — “We take the subway to Ruhleben” (Ruhleben is a district of Berlin) was embedded into 16 different contexts. These include (1) a broad focus condition, (2) narrow focus on ‘U-Bahn’ or (3) narrow focus on ‘Ruhleben’. The target sentence was uttered in question and statement mode as a single-phrase or part of a two-phrase utterance. Besides, a variant with a phrase boundary after ‘U-Bahn’ was produced. Table 4.8 gives an overview of all variants examined. Contexts 10 to 12 represent a sequence of two echo-questions.

For illustration, we give examples for Contexts 8, 13 and 15 in which the target phrase is part of a two-phrase utterance. Narrowly focused items are set in bold face.

**Context 8:** “Wir fahren mit der **U-Bahn** nach Ruhleben, weil es mit dem Bus zu lange dauert.” — “We take the **subway** to Ruhleben, because the bus takes too long.”

**Context 13:** “Wir fahren mit der U-Bahn, nach Ruhleben müssen wir nicht umsteigen.” — “We take the subway, to Ruhleben we don’t need to change trains.”

**Context 15:** “Wir müssen keinen Parkplatz suchen, denn wir fahren mit der **U-Bahn** nach Ruhleben.” — “We don’t need to search for parking space, since we take the **subway** to Ruhleben.”

All variants were uttered by a native speaker of German five times at an average speed of 6.5 syllables per second. The speech data were directly sampled at 16 kHz/16 bit using a PC soundcard. The $F_0$ contours of all utterances were analyzed with the Fujisaki model using the Analysis-by-Synthesis method aiming at reducing the mean square error in the log $F_0$ domain. Segment boundaries were marked auditorily on the speech waveform.
4.4. **FOCUS, SENTENCE MODE AND PHRASE BOUNDARY LOCATION**

Figure 4.8: Example of analysis for a sample of Context 1. The figure displays from top to bottom: the speech waveform, the extracted (plus-signs) and model-based $F_0$ contours (solid line) $F_0$, the syllable-based $z$-score averaged over five repetitions, and the underlying accent commands. The vertical lines denote syllable boundaries.

Figure 4.9: Example of analysis for a sample of Context 2 (narrow focus on ‘U-Bahn’).
4.4.3 Results of Analysis

Figures 4.8 to 4.10 show results of analysis for samples of Contexts 1, 2 and 3. The figures display from top to bottom: the speech waveform, the extracted (plus-signs) and model-based \( F_0 \) contours (solid line), the syllable-based z-score (defined as \( (t_{sy} - t_{\mu}) / t_{\sigma} \) for log durations), and the underlying accent commands. The vertical lines denote syllable boundaries. It must be noted that all syllable-based z-scores drawn in the figures in this section are values averaged over all renditions of the same context.

The Influence of the Focal Condition  The evaluation first concentrates on the effect of focus shift on the word accent syllable durations and tone switch intervals at the potentially accented items ‘neh-men’, ‘U-Bahn’ and ‘Ruh-le-ben’. Table 4.9 gives mean syllable durations \( t_{\mu} \) and tone intervals (in terms of changes of \( A_a \), denoted as ‘\( A_a_{\mu} \)’) for all 16 contexts. Cases where a respective item is narrowly focused are marked by bold face. If we examine the results for the item ‘U-Bahn’ we find that a narrow focus results in a significant boost of the tone switch, as can be seen from the value of \( A_a \) comparing Contexts 2 (Figure 4.9) with Contexts 1 (Figure 4.8) and 3 (Figure 4.10), for instance.

The duration of the accented syllable ‘U’ varies by an average 27 % between deaccented (narrow focus on ‘Ruhleben’) and narrow focus conditions. Independent-samples T-Test shows that these results are all highly significant (\( p < .01 \)). Under broad focus condition, on the average, the syllable ‘U’ is 8 % longer than in the de-accented version, but this result does not prove to be significant (\( p < .08 \)). In other words: for the syllable ‘U’ duration and tone switch interval are highly correlated (\( \rho = .69 \)).

The tone switch interval for the item ‘Ruhleben’ is influenced in a similar way as the one on ‘U-Bahn’ since it is boosted when ‘Ruhleben’ becomes narrowly focused. However, the accent

![Table 4.9: Mean word accent syllable durations and tone switch intervals for the potentially accented syllables ‘neh’, ‘U’ and ‘Ruh’.

<table>
<thead>
<tr>
<th>No.</th>
<th>'neh'</th>
<th>'U'</th>
<th>'Ruh'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( t_{\mu} ) [ms]</td>
<td>( A_a_{\mu} )</td>
<td>( t_{\mu} ) [ms]</td>
</tr>
<tr>
<td>1</td>
<td>200.48</td>
<td>138.20</td>
<td>141.29</td>
</tr>
<tr>
<td>2</td>
<td>207.00</td>
<td>154.60</td>
<td>135.00</td>
</tr>
<tr>
<td>3</td>
<td>198.12</td>
<td>141.17</td>
<td>143.41</td>
</tr>
<tr>
<td>4</td>
<td>206.00</td>
<td>161.69</td>
<td>139.13</td>
</tr>
<tr>
<td>5</td>
<td>200.00</td>
<td>126.12</td>
<td>159.40</td>
</tr>
<tr>
<td>6</td>
<td>210.27</td>
<td>143.36</td>
<td>155.47</td>
</tr>
<tr>
<td>7</td>
<td>206.00</td>
<td>164.83</td>
<td>152.15</td>
</tr>
<tr>
<td>8</td>
<td>191.04</td>
<td>139.05</td>
<td>139.69</td>
</tr>
<tr>
<td>9</td>
<td>194.13</td>
<td>124.21</td>
<td>152.45</td>
</tr>
<tr>
<td>10</td>
<td>177.35</td>
<td>168.52</td>
<td>153.23</td>
</tr>
<tr>
<td>11</td>
<td>205.00</td>
<td>119.12</td>
<td>162.71</td>
</tr>
<tr>
<td>12</td>
<td>169.00</td>
<td>134.44</td>
<td>126.48</td>
</tr>
<tr>
<td>13</td>
<td>191.20</td>
<td>134.29</td>
<td>129.56</td>
</tr>
<tr>
<td>14</td>
<td>187.05</td>
<td>155.68</td>
<td>145.00</td>
</tr>
<tr>
<td>15</td>
<td>168.00</td>
<td>133.08</td>
<td>133.65</td>
</tr>
</tbody>
</table>

Table 4.9: Mean word accent syllable durations and tone switch intervals for the potentially accented syllables ‘neh’, ‘U’ and ‘Ruh’.


4.4. FOCUS, SENTENCE MODE AND PHRASE BOUNDARY LOCATION

Figure 4.10: An example of analysis from Context 3 (narrow focus on ‘Ruhleben’).

Figure 4.11: Syllable z-scores averaged over the three different focal condition BROAD, narrow focus on ‘U-Bahn’ and ‘Ruhleben’, respectively. CNTXT13 denotes the context in which a syllable boundary occurs after ‘U-Bahn’.
syllable ‘Ruh’ does not exhibit any significant duration change ($p > .336$).

The tone switch at the item ‘nehmen’ is clearly reduced under all narrow focus conditions, whereas the duration of the accent syllable ‘neh’ remains largely unaffected.

**Sentence Mode** Since it can be expected that the sentence mode mostly influences the final part of a phrase we examined the duration of the syllable ‘ben’ under all contexts and found that it is significantly longer in statement-final than in question-final position (199 vs. 153 ms, $p < .01$). In Context 13 where it is the third syllable of the second phrase it is further compressed to an average of 123 ms. Statements are generally marked by a negative tone switch at the last accented item in the phrase, whereas questions exhibit a positive tone switch at this item and a question-final rise on the last syllable ‘ben’.

**Phrase Boundary Location** If we compare the results from Context 13 with those from Context 7, both of which represent two-phrase statements under broad focus, the vastest difference found is the lengthening of the syllable ‘Bahn’ in Context 13 (an average 318 ms against 200 ms, $p < .01$). The tone switch on the pre-boundary accent syllable ‘U’ in Context 13 is only slightly higher than that in Context 7 (.44 vs. .36). In order to summarize the results on syllable durations, Figure 4.11 gives averaged syllable-based z-scores for all different focal conditions. In addition, the results from Context 13 are displayed. Whereas the first four syllables “Wir nehmen die” are largely unaffected by focus shift, narrow focus on ‘U-Bahn’ leads to a considerable lengthening of the accent syllable ‘U’-’. In contrast, narrow focus on ‘Ruh-leben’ compresses the pre-focal syllables ‘U’-, ‘Bahn’ und ‘nach’. Comparison with Context 13 shows that the phrase boundary after ‘U-Bahn’ stretches the phrase-final syllable ‘Bahn’, while compressing the second phrase-initial syllables ‘nach’, ‘Ruh’ and ‘le’. This indicates a certain compensatory effect.

4.4.4 Discussion and Conclusions

On the data presented here, it was observed that focus shift influences the $F_0$ contour more strongly and in a more uniform way than syllable durations. This might be explained by the greater freedom in the use of tone switches for coding prominence information, since they can be completely deleted, whereas syllables cannot be compressed below a certain length. Furthermore, it is interesting to note that focus shift influences phrase-medial items in a different way than phrase-final ones as the latter do not show any significant durational changes. As far as the indication of phrase boundaries is concerned, syllable duration obviously plays a more imminent role than modifications of the pre-phrase boundary tone switch.

4.5 Summary of Preliminary Studies

The following main results can be stated in terms of the relationship between the $F_0$ contour and the durational structure of an utterance:

- The fine timing of accent commands can be reliably related to the syllable as a reference. Whereas $T1$ is strongly influenced by the internal structure of the syllable in terms of onset, nucleus and overall syllable duration, $T2$ is timed with respect to the syllable offset. The accent command amplitude $Aa$ does not exhibit a significant influence on the accent command timing.
4.5. SUMMARY OF PRELIMINARY STUDIES

- Perceived prominence is correlated with accent command amplitude $A_a$. This means that $A_a$ reflects the organization within an utterance in terms of highlighting or dimming certain constituents. This is especially noticeable under varying focus conditions.

- Syllable durations are also strongly related with perceived prominence, but they are the predominant feature for lower prominence levels, i.e., in the absence of accent commands. This further motivates the distinction between stress and accent adopted in this thesis, as the latter is mainly connected with $F_0$.

- In terms of variability, syllable durations vary considerably little under different focus (i.e., degree of accentuation) conditions, but their lengthening is connected with phrase boundaries. Whereas syllables can be compressed or lengthened only within a certain range, tone switches can be completely deleted or very strongly boosted.
Chapter 5

An Integrated Approach to Modeling Prosody

Abstract
This chapter describes the design and development of the integrated model of prosody. In this context we shall point out the correspondence between the intonational units defined in MFGI and durational units in the light of a preliminary production experiment. The framework for extending MFGI into an integrated prosodic model is developed by identifying the informational units to be coded and the prosodic representations to be output by the model. In the last part of the chapter the speech database used, and the procedure for training and testing the model are discussed.
5.1 Introduction

As we have seen in the preceding chapters most conventional TTS systems for German like DRESS calculate prosodic parameters sequentially as indicated by Figure 5.1. This means that the \( F_0 \) contour is aligned to the speech signal with respect to the segment durations determined by the duration control and therefore is affected by the imperfections of the latter. Furthermore, the modules for predicting duration and \( F_0 \) are often developed independently and use features derived from different data sources and environments. This approach neglects the fact that the \( F_0 \) contour and the duration contour are influenced by the same linguistic, para-linguistic and non-linguistic information units. Generally speaking, conventional prosody models ignore the fact that the natural speech signal is coherent in the sense that intonation and speech rhythm are co-occurrent and hence strongly correlated as indicated by Figure 5.2. This shortcoming partly explains why synthetic speech is easily identified and rated as being of poor quality. Based on these considerations, the objective of the author is the development of a prosodic model taking into account the coherence between melodic and rhythmic properties of speech.

In other words, the main design rationale of the approach is based on the observed interactions between the prosodic features of speech as shown in the studies discussed in the preceding chapters. The resulting model can be called a parallel or ‘integrated’ approach. In Figure 5.3 the block diagram of such an integrated approach is displayed. In order to build a model one needs to determine which linguistic and phonetic factors exert major influence on the \( F_0 \) contour and syllable durations.

The project of developing and evaluating the integrated model (henceforth IGM) of prosody documented in the current and the following chapter can be subdivided into the following phases:

- Database selection and data extraction (Sections 5.2 and 5.3),
- Statistical analysis of prosodic data (Section 5.4),
- Model development and optimization (Section 5.5),
- Perceptual evaluation (Chapter 6).

5.2 Selecting a Database

A number of freely available speech corpora was considered for developing the IGM. Among these were

- The KIEL corpus (Isolated sentences and stories) [K+97],
- The Bonn Prosodic Database (Isolated sentences and stories) [Heu99],
Figure 5.2: Speech production is a highly coordinated process involving respiration, phonation, and articulation.

Figure 5.3: Block diagram of an integrated prosody model. Duration and $F_0$ contour are calculated in parallel.
5.3. PARAMETER EXTRACTION AND ORGANIZATION

- The VERBMOBIL database (quasi-spontaneous utterances from a scheduling task) [RB94],
- The Stuttgart Radio Corpus (radio news from the Deutschlandfunk recorded in 1995) [Rap98].

Eventually the Stuttgart Radio Corpus was chosen for the following reasons:

- The data is informative real-life material, not typical ‘lab speech’.
- It covers basically unrestricted informative texts produced by four professional newscasts in a neutral manner.
- The subcorpus eventually used for creating the model is reasonably large: 48 minutes of news stories read by a male speaker, of a total of 13151 syllables.
- The other corpora except for the VERBMOBIL corpus are rather heterogeneous and comprise a relatively large number of speakers in different speaking styles, but with relatively little material for each condition.
- The VERBMOBIL corpus which was at first considered, consists of relatively short utterances in a dialogue style and therefore eventually did not appear to be a good basis for deriving prosodic features for a TTS system which in many applications serves as a reading machine.

The Stuttgart corpus contains boundary labels on the phone, syllable and word levels and linguistic annotations such as part-of-speech. Prosodic labels for tones and boundary depth are provided according to the Stuttgart ToBI system [May95].

5.3 Parameter Extraction and Organization

5.3.1 Organizing the Database

Before the information stored in several database files per utterance can be processed, it must first be integrated into a linguistically and phonetically rich structure, as indicated in Figure 5.4. This means that by evaluating the temporal information in the label files, phones are assigned to syllables, syllables assigned to words and words assigned to phrases. On each level of representation, relationships are established, such as the index of an item in a superordinate structure, the number of subordinate elements it contains, its onset and offset times, and hence its duration and further properties. For this purpose, a program was written which extracts information on a specific level of representation (the phone, the syllable etc.) and outputs the requested parameters in a table which then can be processed with a statistics toolkit.

Since the syllable was chosen as the basic temporal unit of the model, Table 5.1 shows a (non-exhaustive) list of parameters that were extracted with respect to the syllable. This list, of course, expanded with the developing requirements of the statistical analysis.

Whereas the temporal alignment of phones, syllables, words and phrases was performed fully automatically, the correct alignment of Fujisaki parameters with the syllables required a ‘fuzzy’ search and a considerable amount of manual post-processing. The following section deals with the extraction procedure and also discusses the post-processing.

5.3.2 Fujisaki Model Parameter Estimation

The direct estimation of parameters for the Fujisaki model from the extracted $F_0$ contour poses problems since its components are superimposed in a particular contour and difficult to be
Figure 5.4: Integrating the information contained in the database into a linguistic structure.
### 5.3. PARAMETER EXTRACTION AND ORGANIZATION

Table 5.1: Parameters on the syllable level extracted from the database.

<table>
<thead>
<tr>
<th>all syllables</th>
<th>accented syllables</th>
<th>phrase-initial syllables</th>
</tr>
</thead>
<tbody>
<tr>
<td>onset time ( t_{on} )</td>
<td>ToBI label</td>
<td>( T'_{0} ) of phrase command</td>
</tr>
<tr>
<td>offset time ( t_{off} )</td>
<td>( A )</td>
<td>( A_p ) of phrase command</td>
</tr>
<tr>
<td>index in superordinate word</td>
<td>( T_1 ) of accent command</td>
<td>syllables in preceding phrase</td>
</tr>
<tr>
<td>index in foot</td>
<td>( T_2 ) of accent command</td>
<td>( \alpha ) of phrase command</td>
</tr>
<tr>
<td>index in superordinate phrase</td>
<td>( T_{1_{dist}} )</td>
<td>beginning of paragraph</td>
</tr>
<tr>
<td>index from preceding phrase command</td>
<td>( T_{2_{dist}} )</td>
<td>beginning of sentence</td>
</tr>
<tr>
<td>p.o.s. of superordinate word</td>
<td>( T_{1_{rel}} )</td>
<td>beginning of phrase</td>
</tr>
<tr>
<td>Break index to the left</td>
<td>( T_{2_{rel}} )</td>
<td>( T_{0_{dist}} ) duration preceding pause</td>
</tr>
<tr>
<td>Break index to the right</td>
<td>last accent in phrase</td>
<td>duration current phrase</td>
</tr>
<tr>
<td>text of superordinate word</td>
<td>orientation</td>
<td></td>
</tr>
<tr>
<td>SAMPA transcription of syllable</td>
<td>intoneme ( \beta )</td>
<td></td>
</tr>
<tr>
<td>accented(y/n)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lexical accent (y/n)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>index of phrase in sentence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>index of sentence in paragraph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPA of syllable onset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPA of syllable nucleus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPA of syllable coda</td>
<td></td>
<td></td>
</tr>
<tr>
<td>duration onset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>duration nucleus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>duration coda</td>
<td></td>
<td></td>
</tr>
<tr>
<td>number phones onset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>number phones nucleus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>number phones coda</td>
<td></td>
<td></td>
</tr>
<tr>
<td>weight as defined in [Loc.94]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>distance from major break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>distance from nearest accent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>first syllable in phrase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>last syllable in phrase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>power rms (intensity)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of phones in syllable</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
inferred directly. Furthermore, determining the appropriate number of model commands underlying a given $F_0$ contour requires a trade-off between fitting accuracy and linguistic meaningfulness. As a consequence, methods for determining model parameters were either limited to short utterances [Geo93, MPH93] or required user interaction flawing the objectiveness of the analysis.

At the initial stage of his D.Eng. research (see [Mix98, p. 66ff.]) the author chose to employ a semi-automatic approach. As a first step linguistic information was derived from the text underlying the utterances as to the locations and intoneme types of potentially accented syllables and the locations of phrase boundaries. These were determined by applying default intonation rules following [SZ82] for initializing the appropriate number of Fujisaki model commands. The resulting command configuration was then optimized locally by approximating the natural $F_0$ contour. The sequence of analysis steps is illustrated in Figure 5.6.

It must be noted that the primary objective in this work was not the ‘blind’ extraction of prosodic cues for speech recognition (as pursued by [Str95], for instance), but the collection of parameters for predicting $F_0$ contours in TTS.

The current section presents a robust multi-stage approach consisting of a quadratic spline smoothing, contour filtering, accent command initialization and a three-pass Analysis-by-Synthesis procedure [Mix00]. It shares certain ideas, such as the contour filtering with earlier works by Strom [Str95].

The $F_0$ values in the Stuttgart Radio Corpus are provided for intervals of 10 ms, along with
Figure 5.6: Procedure of analysis for Fujisaki-model parameter extraction explained in 6 steps: 1) extraction of raw $F_0$ contour, 2) marking of word boundaries, correction of $F_0$ detection errors, marking of reliable section (+ signs: $R(T_0) > 0.9$, □ signs: $R(T_0) \leq 0.9$), 3) from the text: candidates for phrase boundaries and accents, 4) Pre-selection of phrase commands and $Fb$, 5) Pre-selection of accent commands, 6) Numerical Optimization.
frame-wise intensity- and degree-of-vocing-measures. The latter are used for weighting the $F_0$ contour in the final phase of the modeling procedure.

![Graphs](image)

Figure 5.7: Initial Fujisaki model parameter configuration, bottom: phrase and accent commands, center: LFC and resulting phrase component, top: HFC and resulting accent component.

### 5.3.2.1 Quadratic Spline Stylization

Prior to modeling a given $F_0$ contour, two tasks are performed: (1) Intermediate $F_0$ values for unvoiced speech segments and short pauses are interpolated from the extracted $F_0$ contour, (2) Microprosodic variations caused by the influence of individual speech sounds (explosion, frication, etc.) are smoothed out, as the Fujisaki model explicitly deals with macroprosody only. One method successfully applied to the two tasks mentioned is the MOMEL model [HE93] which converts a given $F_0$ contour into a sequence of target points used as a reference for performing a spline interpolation of the contour. It has been shown that MOMEL can be applied regardless of the particular language.

Figure 5.5 (top) shows the initial part of an example utterance displaying the extracted (+ signs) and the spline contours (solid line).

### 5.3.2.2 High-Pass Filtering and Component Separation

In order to separate the accent component from the phrase component and $Fb$, the spline contour is passed through a high-pass filter with a stop frequency at 0.5 Hz. The output of the high-pass (henceforth called 'high frequency contour' or HFC) is subtracted from the spline
contour yielding a ‘low frequency contour’ (LFC) which contains the sum of phrase component and $Fb$. The latter is initially set to the overall minimum of the LFC. Hence, partial contours roughly corresponding to phrase and accent components are determined, as shown in Figure 5.7 (bottom).

### 5.3.2.3 Command Initialization

The initialization procedure makes use of the characteristics of phrase and accent command responses making up phrase and accent components, respectively. As can be seen from Figure 2.36, in a sequence of phrase commands, the onset of a new command is characterized by a local minimum in the phrase component. Consequently, the LFC is searched for local minima, applying a minimum distance threshold of 1 s between consecutive phrase commands. To initialize the magnitude value $Ap$ assigned to each phrase command the part of the LFC after the potential onset time $T0$ of a phrase command is searched for the closest local maximum. $Ap$ is then calculated in proportion to the frequency value found at this point. As responses of several phrase commands may add up in the phrase component, contributions of preceding commands must be taken into account when calculating $Ap$, which is reduced accordingly (see Figure 5.7, center). A full phrase command reset occurs at inter-phrase boundaries accompanied by a longer pause (> 300 ms). The time constant $\alpha$ is initially set to 0.95/s, a value found appropriate after a series of preliminary trials.
To initialize the appropriate number, onset times $T_1$ and offset times $T_2$ of accent commands, the HFC is searched for local minima, whose vicinity ($\pm100$ ms) is scanned for even lower $F_0$ values in order to avoid picking saddle points. Two subsequent local minima each are associated with a new accent command. Since the accent command response requires some time to decay to 0 after $T_2$, $T_2$ is set back to 200 ms before the local minimum. The accent command time constant $\beta$ is set to an initial value of 20/s. To initialize the accent command amplitude $A_a$, the maximum in the HFC between $T_1$ and $T_2$ is determined, and $A_a$ is set in proportion to the frequency value found at this point (see Figure 5.7, top). Accent commands are not continued across major pauses in the speech signal, as is the case for the right-most accent command in Figure 5.7, bottom.

5.3.2.4 Analysis-by-Synthesis

The analysis by synthesis procedure is performed in three steps, in the course of which the initial parameter configuration (Figure 5.7, bottom) is subsequently optimized by applying a hill-climb search for reducing the overall mean-square error in the log $F_0$ domain. Each step terminates when the improvement between subsequent iterations drops below a set threshold. At the first step, phrase and accent components are optimized separately, taking the LFC and HFC, respectively, as the targets. Figure 5.8, panel (1) shows the joint result of this step which already yields a quite close approximation of the spline contour. Next, phrase component, accent component and $Fb$ are optimized jointly, taking the spline contour itself as the target (see Figure 5.8, panel (2) for the resulting approximation). In the final step, the parameter configuration is further fine-tuned by making use of a weighted representation of the extracted original $F_0$ contour. The weighting factor applied is the product of degree of voicing and frame intensity for every $F_0$ value, which favors 'reliable' portions of the contour, Figure 5.8, panels (3) and (4) show the resulting model contour and the underlying model commands. The mean approximation error yielded with the reported approach on the entire speech database amounts to 3.1 % when taking into account all voiced frames, but is considerably lower for reliable parts of the contour, such as stable vowel portions (mean error: 1.7 %). Before longer inter-phrase pauses, the stylization algorithm occasionally levels short rising sections of the $F_0$ contour belonging to boundary tones, an error from which the analysis procedure cannot recover. Similarly, the procedure cannot recover if too few commands were set up in the initialization phase. In some rare cases, a missing phrase command was hence compensated by a very small but long accent command. Superfluous commands can usually be identified by their rather short durations (< 50 ms) and small amplitudes (< 0.1) and are removed between analysis steps.

The database was then examined manually for accent commands which could not be justified by the vicinity of accented syllables or boundary tones. These were deleted and the analysis procedure was resumed with last stage.

5.4 Statistical Analysis of the Prosodic Database

5.4.1 Introduction

In order to develop the IGM a statistical analysis of the prosodic database was performed. This analysis aimed at identifying the most important input factors of the model and their influence on the prosodic output features. Following the consideration that the syllable plays a crucial role in the production as well as the perception of speech (see Section 4.3) it was chosen as the basic rhytmical unit for the IGM. Therefore at first the main factors influencing syllable duration are investigated. This work is documented in the following section.
5.4. STATISTICAL ANALYSIS OF THE PROSODIC DATABASE

Secondly, it is necessary to examine how intonational units in terms of Fujisaki control commands are best 'linked' to the syllabic grid constituting the rhythmical structure of an utterance. And thirdly, the most important input features of the prosodic model influencing the $F_0$ contour have to be identified.

5.4.2 Analysis of Syllable Durations

5.4.2.1 Introduction

Throughout this thesis, the example of syllable durations was the only context in which a comparison of three different prediction approaches was performed, namely

- Feed-forward neural networks (FFNN),
- CARTs,
- Regression models.

For the remaining part of the thesis FFNNs were employed exclusively, namely for the fact that they permit the joint prediction of several prosodic features, whereas CARTs and simple regression models are limited to predicting single parameters. The comparison on syllabic durations of the three approaches was therefore mainly conducted to ensure that FFNNs performed as well as the other methods.

In the following section factors which in earlier works (see Section 2.6.2) were found to influence syllable duration were examined for introduction into the IGM.

5.4.2.2 A Model of Syllable Durations

Motivation The accurate prediction of syllabic durations must be regarded as the centerpiece of the IGM, since

- The accent commands of the Fujisaki model require timing information extracted from the segmental string, and as a first approximation $T_1$ and $T_2$ can be related to the accent syllable duration (see Section 4.2).

- Not all utterance-medial phrase boundaries are accompanied by phrase commands of the Fujisaki model, hence additional features such as pauses and local maxima of the relative syllable duration contour may be more reliable markers of these boundaries.

- Syllable duration is known to be influenced by a number of internal and external factors, the former concerning the syllable structure, such as the total number of phones, the nucleus properties (quantity, place of articulation, the tendency to expand or reduce) and coda property (voicing), the latter determined by the presence / absence of stress, accent and boundaries to the right of the syllable, and positional aspects, such as the distance from a major phrase boundary ($BI \geq 3$), and the position in the foot expressed by the distance from the head. A regression model can help to separate the relative contributions of internal and external factors to the observed duration of a particular syllable, and serve as predictor for syllables not contained in the database.

- Using the regression model as a descriptive tool, the observed duration contour (i.e. $dur_{obs}$, i.e. index of syllable in an utterance) can be decomposed into two contours — one (mainly) influenced by internal factors and one influenced by external factors. Analogous to the distinction between micro and macro prosody in the $F_0$ contour, the influence
of the external factors (mostly linguistic) factors on the duration contour corresponds to
the macroprosodic domain, the internal factors correspond to the microprosodic domain.
Since we are mostly interested in the macroprosody, we need to normalize the duration
contour with respect to the internal structure.

Factors examined for Introduction into the Model of Syllable Duration
The objective of this statistical analysis is the identification of factors which can most appropriately account
for the larger part of variation in syllable duration observed throughout the database. Since
some input factors themselves are correlated, the added descriptive power of these factors is
usually lower than the theoretical sum of the particular contributions. Hence the construction
of the model requires a trade-off between the number of factors introduced and the percentage
of variation that can be explained. In the following the factors examined so far will be discussed.
It is assumed that segmentation inaccuracies are normal-distributed, hence in the ideal case that
all important factors are introduced to the model, the resulting prediction error should also be
normal-distributed.

5.4.2.3 Extrinsic factors

Syllable Strength The word accent syllable is the location of lexical accent in a word
(‘Mauer’ — ‘wall’, ‘gefahren’ — ‘driven’, ‘bequem’ — ‘comfortable’). In content words (verbs,
nouns, adjectives, etc.), by default, these syllables are classified as stressed, in function words
(articles, prepositions, etc.) they are classified as unstressed. If syllables in the database bear
tone accent labels, they are classified as stressed and accented. All remaining syllables are clas-
sified as unstressed and unaccented. Hence the following categories result for syllable strength:
4. Statistical Analysis of the Prosodic Database

![Graph showing the dependency of syllable duration on BI to the right of the syllable.]

Figure 5.10: Dependency of syllable duration on BI to the right of the syllable.

0  unstressed and unaccented syllables in function and content words,
   unaccented word accent syllables in function words
1  stressed but unaccented (unaccented word accents syllables in
   content words)
2  stressed and accented

Figure 5.9 displays the mean syllable duration depending on the syllable strength.

**Strength of Syntactic Boundary (Syntactic Break Index)**  Boundary strength takes into account the strength of the syntactic boundary to the right of the particular syllable. It should be noted that boundary strength is a mixed factor based on (1) the (theoretical) syntactic boundary strength and (2) prosodic boundaries as observed in the realization of a particular utterance. A larger part of higher level boundaries can very consistently be predicted from punctuation marks, such as commas and periods, the former being assigned a *Syntactic BI* of 3, the latter of 4. Statistical evaluation shows that the speaker marks these boundaries very reliably (close to 100 %) by the introduction of pauses, for instance. In the absence of punctuation marks, additional boundaries are inserted by the speaker which are also usually syntactically motivated, but sometimes also triggered by self-repair. As the marking of BIs 3 and 4 in the ToBI tier is not always consistent, a differentiation between $BI = 3$ for utterance-medial and $BI = 4$ for utterance-final breaks was enforced, hence reducing all original utterance-medial BIs of 4 to $BI = 3$. The inter-word $BI$ of 1 in the ToBI tier was set to 0 in the case of clitic boundaries (article + noun, for instance).
0 Intra-word syllable boundary, inter-word boundary clitic
1 Inter-word boundary non-clitic
2 Intra-phrase prosodic boundary (corresponds to ToBI \textit{BI} = 2)
3 Utterance-medial phrase boundary (‘;’ boundaries, for instance)
4 Utterance-final phrase boundary (‘.’-boundary)

Figure 5.10 displays the mean syllable duration depending on the strength of the phrase boundary. It can be seen that syllables preceding boundaries of index 4 are slightly shorter than those preceding boundaries of index 3. This effect might be explained by syntactic factors, such as the frequency of infinite verbs at the tail of a sentence ending in the suffix [\texttt{-@n}] which is prone to reduction.

**Distance from Major Phrase Boundary** This factor denotes the distance in syllables of a particular syllable from the nearest following major phrase boundary (Syntactic\textit{BI} of 3 or 4). Figure 5.11 above shows the mean syllable duration depending on the distance from the nearest major phrase boundary as a function of the syllable strength. It can be seen that the lengthening effect mainly concerns the syllable immediately preceding the boundary and, to a much smaller degree the penultimate. This pattern prevails irrespective of syllable strength. At least for the material there cannot be documented any rallentando effect over the intonation phrase as reported by Dankovičová [Dan99], but rather a local effect at the phrase boundary.

**Position in Foot** This factor denotes the distance in syllables from the head of the superordinate foot (accented syllable) of a particular syllable, hence the head proper receives a value of 0. In the case of incomplete feet, syllables are assigned a negative value denoting the distance from the next following accented syllable. Figure 5.12 displays the mean syllable duration as a function of the index of a syllable in a foot. It can be seen that the syllable duration generally decreases with the distance from the head.

**5.4.2.4 Intrinsic Factors**

As mentioned above, these factors concern the internal structure of the syllable.

**Number of Phones in Syllable.** Obviously the duration of a syllable strongly depends on the number of phones which it contains. Due to compensatory effects, however, the individual phone durations do not exactly add up. Furthermore, the phones vary as to their inherent durations. Therefore the number of phones can only be a first approximation where the individual duration means of the phone classes are not known.

**Schwa** Syllables containing a schwa as the nuclear vowel are prone to reduction.

**Place** The place of articulation of the nuclear vowel (back/front).

\begin{verbatim}
Coding: 0 @, 9, a, @, U, i, O, OY, o:
        1 I, a, E, E, y, 2:
\end{verbatim}

**Short** The long/short distinction of nuclear vowels.

\begin{verbatim}
Coding: 0 i, E, e, 2, a, @, OY, y:
        1 @, I, E, a, O, U, y
\end{verbatim}

**Long** The tendency of certain vowels to be lengthened more easily than others.

\begin{verbatim}
Coding: 0 I, U, Y, @, e, i, u, @, y, 2:
        1 a, E, O, 9, a, E, a, aU, OY
\end{verbatim}
5.4. **Statistical Analysis of the Prosodic Database**

![Distance from major boundary in syllables](image1)

**Distance from major boundary in syllables**

Figure 5.11: Dependency of syllable duration on distance from major prosodic boundary.

![Syllable index in foot](image2)

**Syllable index in foot**

Figure 5.12: Dependency of syllable duration on position in foot.
Table 5.2: Correlation between Syllable Duration and Factors examined.

<table>
<thead>
<tr>
<th>Factor $x$</th>
<th>$\rho(duration, x)$</th>
<th>Variation Explained [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syllable Strength</td>
<td>.35</td>
<td>12.1</td>
</tr>
<tr>
<td>Strength of Syntactic Boundary (SyntacticBreak Index)</td>
<td>.46</td>
<td>21.5</td>
</tr>
<tr>
<td>Distance from Major Phrase Boundary</td>
<td>$- .25$</td>
<td>6.1</td>
</tr>
<tr>
<td>Position in Foot</td>
<td>$- .09$</td>
<td>8.0</td>
</tr>
<tr>
<td>Number of Phones in Syllable</td>
<td>.57</td>
<td>32.0</td>
</tr>
<tr>
<td>Schwa</td>
<td>$- .19$</td>
<td>3.6</td>
</tr>
<tr>
<td>Place</td>
<td>$- .10$</td>
<td>1.0</td>
</tr>
<tr>
<td>Nuclear Vowel Quantity ‘short’</td>
<td>$- .04$</td>
<td>2.0</td>
</tr>
<tr>
<td>Long</td>
<td>.24</td>
<td>5.6</td>
</tr>
<tr>
<td>Coda Voicing</td>
<td>.11</td>
<td>1.2</td>
</tr>
<tr>
<td>Consonant Following Nuclear Vowel Voiced</td>
<td>.15</td>
<td>2.2</td>
</tr>
</tbody>
</table>

**Coda voicing** The property of the syllable coda to be voiced, unvoiced or mixed.

1. coda unvoiced  
2. majority of phones in coda unvoiced  
3. majority of phones in coda voiced  
4. coda voiced

**Consonant Following Nuclear Vowel Voiced** The distinction whether or not the consonant immediately following the vowel is voiced.

5.4.2.5 Correlation between Syllable Duration and Factors examined

Most of the results for the correlation between syllable duration and several factors shown in Table 5.2 are in line with earlier studies on syllable duration. They confirm that

- Stressed syllables are usually longer than unstressed ones, and accented syllables longer than unaccented ones
- Pre-boundary syllables are lengthened, with the effect decreasing with the distance of a syllable from the boundary
- Syllables in the head of a foot which per se are accented or at least stressed, are generally longer than foot-medial and final ones
- Syllable duration increases with the number of phones in a syllable
- Syllables containing a schwa as the nuclear vowel are prone to reduction
- Syllables containing a short vowel as the nucleus are shorter
- Syllables containing a back vowel are slightly shorter than those with a central or low vowel
- Syllables with the vowel tendency ‘long’ are lengthened
- Coda voicing slightly increases the syllable duration
Table 5.3: Adding factors to a regression model of syllabic duration.

<table>
<thead>
<tr>
<th>Factor $x$ added</th>
<th>Variation explained by $x$ alone %</th>
<th>Total variation explained %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Phones in Syllable</td>
<td>32.0</td>
<td>32.0</td>
</tr>
<tr>
<td>Strength of Syntactic Boundary</td>
<td>21.5</td>
<td>45.1</td>
</tr>
<tr>
<td>Syllable Strength</td>
<td>12.1</td>
<td>55.1</td>
</tr>
<tr>
<td>Schwa</td>
<td>3.6</td>
<td>59.8</td>
</tr>
<tr>
<td>Place</td>
<td>1.0</td>
<td>61.7</td>
</tr>
<tr>
<td>Long</td>
<td>5.6</td>
<td>62.1</td>
</tr>
<tr>
<td>Short</td>
<td>0.2</td>
<td>62.5</td>
</tr>
<tr>
<td>Distance from Major Phrase Boundary</td>
<td>6.1</td>
<td>62.7</td>
</tr>
<tr>
<td>Consonant Following Nuclear Vowel Voiced</td>
<td>2.2</td>
<td>62.8</td>
</tr>
</tbody>
</table>

The rightmost column indicates the percentage of variation in syllable duration explained by a regression model considering factor $x$ alone (formula: $\text{Dur}_{x} = \text{Constant} + \beta \times x$).

As a first approximation, regression models are calculated which take into account those factors with the most descriptive power, by successively adding factors in descending order of importance and determining the percentage of variation explained. It can be observed that the order of importance of factors is modified as a result of the regression analysis, as some factors, such as 'Strength of Syntactic Boundary' and 'Distance from Major Phrase Boundary', are strongly correlated, hence adding the latter to the model yields little overall improvement (see Table 5.3).

As can be seen from the figures, the first four factors account for about 60% of variation of the syllable duration observed, and adding five more factors to the model only yields another 3%. As a conclusion we hypothesize the following:

1. Major factors are still missing in the model.

2. The way factors are subdivided into categories is inappropriate (important subcategories have been merged in the same category).

3. Using a linear regression model is inappropriate.

4. Predicting linear durations is inappropriate.

1. **Major factors are still missing in the model**. A main factor not captured by the model is the segmentation error incurred by the automatic alignment tool used for the alignment of phone boundaries, since these are caused by the determination of syllable onset, offset and duration. The vast majority of segment boundaries in the database is labeled with a resolution of 10 ms. If we relate the expected segmentation error to the mean $F_0$ for the speaker (95 Hz) this gives a mean $T_0$ of about 10.5 ms.

   As the details of the alignment procedure are unknown it shall be assumed that the segmentation error in the ideal case can be treated as normally distributed noise. Hence, the residual of the regression model should be expected to be nearly normally distributed as well.
2. The way factors are subdivided into categories is inappropriate (important subcategories have been merged in the same category) Compared with the task of predicting phone durations, due to the large number of legal syllable structures of German, a finer resolution of the factor syllable structure (in terms of its SAMPA transcription, for instance), is not feasible, as tokens in the database are not available for all contexts of syllable boundary, strength and position. A possibility would be the clustering of speech segments with similar durational properties as proposed in [ZK98].

3. Using a linear regression model is inappropriate The equation of the linear regression model has the following form:

\[ Dur_{est} = \text{Constant} + \beta_1 \times \text{factor}_1 + \beta_2 \times \text{factor}_2 + \ldots + \beta_N \times \text{factor}_N \]

and

\[ Dur_{obs} = Dur_{est} + \text{error} \]

with \( Dur_{est} \) being the duration estimate for a particular syllable, and \( Dur_{obs} \) the syllable duration observed, \( \text{error} \) the estimation error or residual, \( \beta_1, \beta_2, \ldots, \beta_N \) the regression coefficients, and \( \text{factor}_1, \text{factor}_2, \ldots, \text{factor}_N \) the factors introduced into the model.

Most current models for predicting syllable duration are regression (‘sums of factors’) models or at least regression-based models such as CART trees which aim at the reduction of the overall standard deviation observed. The current regression model is appropriate as to the direction of changes in syllable duration caused by the modification of factor value (stronger syllables are longer, as well as syllables with a larger number of phones, etc.). The model is also appropriate as to the existence of a constant term which can be interpreted as a certain minimum duration below which a syllable cannot be compressed.

It cannot be excluded, however, that

- a multiplicative model of the form

\[ Dur_{est} = Dur_{min} + Dur_{inh} \times \beta_1(\text{factor}_1) \times \beta_2(\text{factor}_2) \times \ldots \times \beta_N(\text{factor}_N) \]

will be more appropriate (see, for instance, [MV96])

- single factors are quadratically, cubically or exponentially related with the duration change they produce (as is known for the number of phones in a syllable which is not exactly linearly related to the syllable duration)

- factors are correlated in a manner that an increase in one factor decreases the effect of another, for instance, i.e. \( \beta_i = f(\text{factor}_j) \)

4. Predicting linear durations is inappropriate The z-score [Cam92] which denotes the relative duration of a phone within a particular database uses log durations and their means and standard deviation which might indicate that the regression model should rather be applied to log durations. However, on the current data using the same factors for predicting log duration yields a comparable 59.0% of variation explained.

Since, however, the ultimate aim of this section was to examine the relative contributions of various factors to syllable duration, the regression model was used as a convenient tool, and optimization issues do not concern us at this moment.

5.4.2.6 Decomposing the Duration Contour

As was mentioned above, the regression model can be applied to separate the contribution of external and internal factors to the duration of a particular syllable. These factors could also be described as linguistic and phonetic factors, the former concerning the syntax of a sentence
and the morphological structure and lexical property of the words in it, the latter concerning the individual syllable structure in terms of its constituent phones. In theory, external and internal factors shall be thought of being mutually independent, which is supported by the fact that adding the external factors syllable and boundary strength to the internal factor number of phones yields a considerable gain in descriptive power of the regression model. In practice, however, syllables of frequent function words, for instance, will always be found to be unaccented (der, die, das, ...), hence introducing a correlation between syllable strength and syllable structure. The same can be observed with syllables containing a schwa as the nucleus, as these are by default unaccented. Despite these considerations, the attempt will be made to use the regression model for separating internal and external factors and decompose the observed duration contour into their contributions. The constant in the equation is treated as part of the internal duration as it is a property of the syllable, and also helps to avoid negative durations in this part.

\[
\text{Dur}_{obs} = \text{Dur}_{est} + \text{error} = \text{Dur}_{int,est} + \text{Dur}_{ext,est} + \text{error}
\]

- \(\text{Dur}_{obs}\): observed syllable duration
- \(\text{Dur}_{est}\): estimated syllable duration
- \(\text{error}\): estimation error
- \(\text{Dur}_{int,est}\): estimated durational contribution of intrinsic factors
- \(\text{Dur}_{ext,est}\): estimated durational contribution of extrinsic factors
If we simply calculate $Dur_{\text{int}}$ and $Dur_{\text{ext}}$ for a particular syllable from the factor values known (number of phones, syllable strength, etc.), the estimated total duration of the syllable will usually not be equal to the observed syllable duration, due to the estimation error of the regression model. Hence the requirement of decomposition as stated above is not met. If we, however, consider that the segmentation error adds noise to both $Dur_{\text{int,est}}$ and $Dur_{\text{ext,est}}$ we can rewrite the above equation in the following form:

$$Dur_{\text{obs}} = error_{\text{factor}} \times Dur_{\text{ext}} = error_{\text{factor}} \times (Dur_{\text{int,est}} + Dur_{\text{ext,est}})$$ (5.2)

and yield a factor for scaling $Dur_{\text{int,est}}$ and $Dur_{\text{ext,est}}$ to sum up to the observed syllable duration. Hence ‘observed’ external and internal duration contours can be calculated in the following form:

$$Dur_{\text{int,obs}} = error_{\text{factor}} \times Dur_{\text{int,est}} Dur_{\text{ext,obs}} = error_{\text{factor}} \times Dur_{\text{ext,est}}$$ (5.3)

Figure 5.13 shows an example of decomposition of the observed duration contour into partial duration contours reflecting the influence of intrinsic and extrinsic factors for the utterance “In der bosnischen Moslem-Enklave Bihać | gingen die Kämpfe zwischen den Regierungstruppen und serbischen Verbänden | auch heute früh weiter |” (‘|’ denoting accented syllables, ’|’ denoting major phrase boundaries). Peaks in $Dur_{\text{ext,obs}}$ reflect accents and higher level boundaries, peaks in $Dur_{\text{int,obs}}$ mark syllables of complex structure, for instance. All contours have been slightly smoothed.

**Analysis of Estimation Error** As mentioned above, if we assume that all main factors have been introduced to the regression model, the remaining prediction error, $Dur_{\text{obs}} - Dur_{\text{est}}$, the
residual, should be mainly due to segmentation inaccuracies and ideally normal-distributed. Figure 5.14 shows a histogram of the estimation error and the corresponding normal curve. The histogram exhibits a slightly left sided maximum, indicating that a considerable number of syllables is estimated slightly longer than actually observed.

The mean relative estimation error $\text{abs}(\text{error}/\text{Dur}_{\text{obs}})$ amounts to 22.4% at a standard deviation of 21.1%. The error is slightly lower for accented syllables (18.3%) than unstressed (24.3%), and lower for syllables with a larger number of phones ($N = 1$, 28.7%, $N = 3$, 21.5%, $N = 5$, 13.3%). These effects, however, are introduced by relating error to $\text{Dur}_{\text{obs}} \approx \text{Dur}_{\text{est}} = f(\text{strength}, n_{\text{phones}}...)$, the error per se is uncorrelated with any of the factors in the regression model, as cross-correlation analysis confirms.

As a favorable property, the regression model exhibits a smaller relative error in the predicted durations of accented syllables and syllables preceding a major boundary which serve as important acoustic cues for structuring an utterance. Little can be said, however, as to the accuracy in terms of perceived naturalness provided, which can only be assessed by means of perception experiments.

**Intra-Speaker-Variation** In order to assess the magnitude of the prediction error incurred by the regression model, we examined the intra-speaker variation in syllable duration. Several passages of text are recurrent in the database, as the same news was read at different hours on the same day. On the example of a passage of 146 syllables that was repeated five times, we calculated the difference between actual and mean duration for each syllable and found a mean deviation of 7.4% with a standard deviation of 7.2%. This is still only about 1/3 of the error of the regression model, but at an average of 20 ms it is considerably larger than the labeling resolution of 10 ms.

![Figure 5.15: Example of display of complete prosodic information. From top to bottom: Speech waveform, extracted and model-generated $F_0$ contours, the extrinsic duration contour in terms of the syllabic $z$-score drawn as horizontal lines of the length of the respective syllable, the ToBI tier, the text of the utterance, the underlying phrase and accent commands.](image)

**Reading Speed** In order to examine if reading speed was missing as a factor in the regression model, the local speech rate in syllables per second was calculated for each syllable at window sizes of 5, 10, 15 and 20 syllables, with the current syllable being the last in the window. For the 5-syllable windows an additional 4.5% of variation was explained when adding the factor to the regression model, which dropped to only 1% for the 20-syllable window. As longer windows are
more likely to overlap the boundary between different sentences which might not belong to the same passage of text, the 5-syllable windows seems optimal. On the other hand, the influence of the current syllable is rather high.

In order to visualize examples from the database with the complete set of underlying prosodic parameters the display used in Figure 5.15 is employed: The figure displays from top to bottom: the speech waveform, the extracted and model-generated $F_0$ contours, the extrinsic duration contour in terms of the syllabic z-score drawn as horizontal lines of the length of the respective syllable, the ToBI tier, the text of the utterance, the underlying phrase and accent commands. The z-score is calculated by clustering syllables with equal intrinsic properties (number of phones, distinction schwa/non-schwa). The duration contour very well indicates above average durations in accented and pre-boundary syllables. The slight hesitation at the beginning of the sentence (which is located at the beginning of a news story) that results in an unusual lengthening of the article 'die' is also very well captured by the duration contour.

In should be noted that in this kind of display syllable durations are treated in the log duration domain, i.e. multiplicatively, as the extrinsic factors influence the intrinsic duration of the syllable by a certain factor. When calculating the intrinsic duration of the syllable, however, the individual linear phone durations are added as will be shown in Section 5.5.

### 5.4.3 Analysis of Fujisaki Control Parameters

#### 5.4.3.1 Introduction

Following the discussion on syllabic durations in the preceding chapter, the current chapter will examine which factors most influence the input parameters of the Fujisaki model. In this context the problem of aligning the model commands with the syllables is also addressed.

![Figure 5.16: Example of tone switch labeling in a statement.](image)

According to the conventions developed in [Mix98], tone switches were labeled as shown in Figure 5.16.

All the prosodic information was arranged in a syllabic grid which was created by temporal alignment of the extracted Fujisaki parameters with the syllables. Figure 5.17 shows an excerpt from the syllable-based table in which all syllabic parameters were arranged.
5.4. STATISTICAL ANALYSIS OF THE PROSODIC DATABASE

5.4.3.2 Aa, Syllable Duration and Intensity

Since Aa as well as the syllable duration are linked with perceived prominence we examined the correlation between these parameters. If we consider all syllables in the corpus we yield a correlation .42 which drops to .28 when we only include accented syllables in the analysis. A scatter plot for this condition is displayed in Figure 5.18. The relatively weak correlation can be explained by the fact that other influences than syllable strength (syllable structure, final lengthening) influence the syllable duration.

The correlation between Aa and syllable intensity is even lower ($\rho = .13$).

5.4.3.3 Fujisaki Model Command Alignment

Since the syllable presents the basic temporal unit of the model, we have to relate the absolute timing of the Fujisaki model commands to the syllabic timing. Hence we calculated the following relative timing parameters:

$T_{0\text{dist}} = t_{on} - T$

$T_{1\text{dist}} = T_{1} - t_{on}$

$T_{2\text{dist}} = T_{2} - t_{off}$

$T_{1\text{rel}} = (T_{1} - t_{on})/(t_{off} - t_{on})$

$T_{2\text{rel}} = (T_{2} - t_{on})/(t_{off} - t_{on})$

As can be seen from the formulations, $T_{1\text{dist}}$ and $T_{2\text{dist}}$ relate the accent command with respect to onset and offset time of the syllable, $T_{1\text{rel}}$ and $T_{2\text{rel}}$ render the timing with respect to the syllable duration. $T_{0\text{dist}}$ expresses the timing of the phrase command preceding a phrase by the distance between the segmental onset of the phrase (i.e. the onset of the first syllable) and T0.
5.4.3.4 General Observations.

Table 5.4 gives mean values for reading speed, $Ap$, the number of phrases (corresponding to the total number of phrase commands found), and $Aa$ for the $N\downarrow$- and $L\downarrow$-intonemes.

The amplitude parameters $Ap$ and $Aa$ vary slightly from speaker to speaker. The following sections will be concerned exclusively with data by male speaker 1 who had spoken by far the largest part of the corpus.

5.4.3.5 Phrasing.

Locations (Mapping of Syntactic Phrases on Intonation Phrases). Analysis shows that the onset of a new paragraph or sentence ($BI$ 4 boundaries) is invariably preceded by a

Table 5.4: Mean values for reading speed and Fujisaki amplitude parameters of the four news speakers.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Fb [Hz]</th>
<th>Speed $S_n$ [Syll/s]</th>
<th>$Ap$ $\mu/\sigma$</th>
<th>$Aa$ $N\downarrow$ $\mu/\sigma$</th>
<th>$Aa$ $L\downarrow$ $\mu/\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>male 1</td>
<td>70.2</td>
<td>7.13</td>
<td>.31/.14</td>
<td>.32/.15</td>
<td>.31/.14</td>
</tr>
<tr>
<td>male 2</td>
<td>76.3</td>
<td>6.87</td>
<td>.41/.18</td>
<td>.31/.13</td>
<td>.21/.05</td>
</tr>
<tr>
<td>female 1</td>
<td>132.5</td>
<td>6.65</td>
<td>.30/.13</td>
<td>.30/.14</td>
<td>.18/.05</td>
</tr>
<tr>
<td>female 2</td>
<td>141.2</td>
<td>6.87</td>
<td>.26/.12</td>
<td>.33/.16</td>
<td>.34/.16</td>
</tr>
</tbody>
</table>
phrase command whereas only about 68 % of BI 3 (intra-sentence) boundaries can be associated with phrase commands. This figure rises slightly in the case of ‘comma’ boundaries (71 %). For intra-sentence boundaries, the timing of phrase commands is more loosely related to the segmental onset of the underlying phrases. The following paragraph renders the text of a news story with the location of extracted phrase commands indicated by ↑, and break index 3 and 4 boundaries given in brackets. Due to their finite time constant phrase commands generally occur within an average distance of 300 ms of the segmental onset of a phrase.


Figure 5.19: Example of problematic assignment of phrase command to segmental phrase.

The example shows that most major syntactic phrases coincide with phrase commands, but in some cases of intra-sentence boundaries the phrase commands occur well into the segmental start of a prosodic phrase (“sagte der parlamentarische...”). This portion of the $F_0$ contour is displayed in Figure 5.19. As the $F_0$ contour suggests, the underlying phrase command does not occur before the phrase onset, but towards the end of the word ‘sagte’.
**Duration of Prosodic Phrases.** About 80% of the phrase command onsets occur within 3.4 s after the preceding command. Figure 5.20 gives the cumulative percentage for the number of syllables of the preceding phrase. About 80% consist of 13 or less syllables. An interesting detail is the fact that utterance-final phrases are generally longer ($\mu = 3.24$ s) than utterance-initial ($\mu = 2.72$ s) or medial ones. This may be explained by the observation that the $F_0$ contour generally reaches its lowest point after the final $1_1$-intoneme where the vocal folds relax and no need is given for phrasing and readjusting the declination line. The phenomenon of a rather low offset value of $F_0$ may well be specific for German, as other languages are reported to behave differently [SZ82, p. 77].

**Table 5.5:** Correlation coefficient $\rho$ between $Ap$ and some parameters.

<table>
<thead>
<tr>
<th>parameter $p$</th>
<th>$\rho(\text{Ap}, p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>boundary depth</td>
<td>.70</td>
</tr>
<tr>
<td>index of phrase in sentence</td>
<td>-.51</td>
</tr>
<tr>
<td>duration of preceding phrase</td>
<td>.32</td>
</tr>
<tr>
<td>duration of current phrase</td>
<td>.11</td>
</tr>
<tr>
<td>$Ap$ of preceding phrase command</td>
<td>-.18</td>
</tr>
</tbody>
</table>

**Phrase Command Magnitude $Ap$.** $Ap$ is the parameter which describes the degree of readjustment of the declination line for the global $F_0$ contour. An important problem connected with
5.4. **Statistical Analysis of the Prosodic Database**

![Graph showing scatter plot of Ap and the number of syllables in preceding phrase.](image)

Figure 5.21: Scatter plot of Ap and the number of syllables in preceding phrase.

Ap is whether the speaker adjusts the declination line depending on the length of the sentence (s)he is going to produce — starting on a higher level for longer utterances — which would be a proof for some kind of utterance pre-planning.

In Table 5.5 we see the correlation coefficients for Ap with the current phrase duration, the duration of the preceding phrase (in seconds or syllables), and the phrase count from head or tail of the utterance. Since the correlation with the duration of the preceding phrase is clearly higher than the respective ratings for the current phrase, we do not find enough evidence for an utterance pre-planning as discussed above. The degree of readjustment of the declination line is more influenced by the time elapsed since the preceding phrase command, which means that it becomes the stronger the lower the F0 contour drops. The correlation with the phrase durations expressed in syllables is smaller, since the segmental string belonging to a phrase often crosses the onset point of the following phrase command, slightly weakening the relationship between.

In Figure 5.21 we see a scatter plot of Ap against the number of syllables in the preceding phrase.

The mean phrase command magnitude for intra-sentence boundaries, inter-sentence-boundaries and paragraph onsets amounts to .8, 1.68, and 2.28 respectively, which shows that Ap is a good correlate of boundary strength. As a consequence the correlation between Ap and the break index to the left of the first syllable in a phrase is relatively high (.70). The negative correlation between Ap and the index of the current phrase indicates that Ap is reduced for phrases later in the utterance which means a reduction of the F0 range in the phrase component.

\[ T_{0, \text{dist}} \]  \[ T_{0, \text{dist}} \] denotes the distance between the phrase command onset and the segmental onset of the respective phrase. For utterance-initial phrases we find nearly a normal distribution
Table 5.6: A selection of parts-of-speech with occurrence, frequency of accentuation and average $\mu$ in the accented case.

<table>
<thead>
<tr>
<th>Part-of-Speech</th>
<th>Occurrence</th>
<th>Accented $%$</th>
<th>Mean $\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>1262</td>
<td>75.8</td>
<td>.28</td>
</tr>
<tr>
<td>Names</td>
<td>311</td>
<td>78.4</td>
<td>.32</td>
</tr>
<tr>
<td>Adjectives conjugated</td>
<td>333</td>
<td>71.6</td>
<td>.25</td>
</tr>
<tr>
<td>Adjectives non-conjugated</td>
<td>97</td>
<td>85.7</td>
<td>.28</td>
</tr>
<tr>
<td>Past participle of full verbs</td>
<td>172</td>
<td>77.3</td>
<td>.29</td>
</tr>
<tr>
<td>Finite full verbs</td>
<td>227</td>
<td>42.7</td>
<td>.30</td>
</tr>
<tr>
<td>Adverbs</td>
<td>279</td>
<td>41.9</td>
<td>.29</td>
</tr>
<tr>
<td>Conjunctions</td>
<td>115</td>
<td>2.6</td>
<td></td>
</tr>
<tr>
<td>Finite auxiliary verb</td>
<td>219</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>Possessive pronouns</td>
<td>65</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>Personal Pronouns</td>
<td>83</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>Articles</td>
<td>804</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Prepositions</td>
<td>621</td>
<td>2.0</td>
<td></td>
</tr>
</tbody>
</table>

around a mean of 490 ms ($\sigma = 370$ ms). In the non-initial case the distribution still spreads around a similar mean value of 406 ms, but is smeared over a larger region ($\sigma = 613$ ms). This reflects the looser connection between phrase commands and phrases in intra-sentence position.

5.4.3.6 Accentuation

Accent Command Amplitude $\mu$. The accent commands from the automatic Fujisaki parameter estimation were contoured with respect to their alignment with syllables bearing the lexical accent. Accent commands which could not be motivated by either accented syllables or boundary tones were removed. As explained in Section 5.3.2, the analysis of accents is a two-sided process. From the text underlying an utterance and by application of accentuation rules [SZ82], the default accent locations are determined and compared with the actual performance structure. At least for the news texts concerned in this study there is a relatively good agreement between predicted and realized accents as will be shown on the following example of a news story (predicted accents are marked with $\text{^P}$, accent command amplitude of commands from the parameter extraction is given in brackets):

"Der SP\textsuperscript{P}(0.15)D-Fraktionsvorsitzende $\text{^P}(0.42)$Scharping wird den $\text{^P}(0.07)$Bundestagsabgeordneten seiner Part\textsuperscript{P}(0.46)tei am $\text{^P}(0.27)$Nachmittag in $\text{^P}(0.27)$Bonn die Ver $\text{^P}(0.30)$trainsfrage stellen. (0.27)Dieser $\text{^P}(0.55)$Schritt sei $\text{^P}(0.21)$notwendig gewor(B:0.51)den, da $\text{^P}(0.28)$Scharping durch seine $\text{^P}$Abwahl als Part\textsuperscript{P}(0.12)teichet be\textsuperscript{P}(0.21)schadigt worden sei, sagte der parlamen\textsuperscript{P}tärische Ge\textsuperscript{P}(0.28)schäftsfrührer der SP\textsuperscript{P}(0.19)D-Fraktion $\text{^P}(0.36)$Struck der Deutschen $\text{^P}(0.17)$Presseagentur. An der $\text{^P}(0.36)$Sitzung wird auch der $\text{^P}(0.12)$neue Par\textsuperscript{P}tvorsitzende $\text{^P}(0.14)$Lafontaine teilnehmen. Nach $\text{^P}$Angaben der $\text{^P}(0.30)$Blickzeitung hat der sarländische Mi\textsuperscript{P}(0.25)isterprü\textsuperscript{P}(0.48)ent unter-$\text{^P}(0.33)$dessen ein Zehn-(0.28)Punkte-Pro(0.43)gramm zum $\text{^P}(0.19)$Kampf gegen die Er$\text{^P}(0.20)$werbslosigkeit (P:0.24)vorgelegt. $\text{^P}(0.47)$Darin sind unter (0.39)anderem kürzere $\text{^P}(0.46)$Arbeitszeiten, fre\textsuperscript{P}(0.29)ilhöhere Ta$\text{^P}(0.28)$rvertr\textsuperscript{P}(B:0.37)ge und längere Ma $\text{^P}(0.28)$schinenlaufzeiten (P:0.21)vorgesehen. Zu$\text{^P}(0.31)$dem sollen $\text{^P}(0.08)$Überstunden $\text{^P}(0.10)$nur noch in $\text{^P}(0.23)$Freizeit abgegol-
5.4. STATISTICAL ANALYSIS OF THE PROSODIC DATABASE

Figure 5.22: Example of post-nuclear accent on ‘vorgesehen’.

(B.0.27)ten und die Lohnnebenkosten ge(0.22)senkt werden.”

From this example we also see two instances of accent commands which belong to boundary tones (marked by ‘B’). Another class of accent commands unaccounted for by the accentuation rules, are those assigned to the lexical stress syllables of verbs following the noun bearing the sentence accent (sometimes called ‘post-nuclear’ accents), marked by ‘P’. Although they are perceptually much weaker than the preceding falling accent (H-down-intoneme) that signals the sentence mode, they exhibit comparable accent amplitudes (see Figure 5.22 for one of the examples). These are intonational events which may not be linguistically motivated, but form part of the phonetic performance structure. For the scope of the current analysis, however, they are not taken into account. Nevertheless they may be worthwhile modeling in a Text-to-Speech system.

In Table 5.6 the frequency of accentuation and the mean $Aa$ is listed for a selection of parts-of-speech. As we can see, nouns and adjectives are accented in over three quarters of cases, but most functions words such as articles very rarely. Verbs and adverbs assume a middle position. Although the figures suggest certain differences in the mean $Aa$ of nouns and proper names, for instance, these values are not significantly different. Hence, the part-of-speech is a strong factor as far as the probability of accentuation is concerned. For accented cases, however, the resulting $Aa$ cannot be reliably predicted from the part-of-speech.

$Aa$, like $Ap$, is reduced for phrases later in the utterance. It is negatively correlated with the index of the accent syllable in a sentence ($\rho = -0.12$). The correlation is much smaller than for $Ap$ though, because $Aa$ is strongly influenced by linguistic factors: Varying focal conditions, for instance, may result in an accent with lower $Aa$ followed by a more prominent one.

The type of intoneme was found to have a relatively strong influence on $Aa$: N-intonemes pre-
ceding an intra-sentence phrase boundary exhibit higher $Aa$ than others (mean of $Aa$ 0.34 against 0.25). The type of accent (non-terminal phrase-final, non-terminal phrase-medial, declarative final) is therefore the most important predictor factor for $Aa$ ($\rho = .29$)

Table 5.7: Relative timing for the different intoneme types.

<table>
<thead>
<tr>
<th>intoneme</th>
<th>$T1_{dist}$ mean/s.d. [ms]</th>
<th>$T2_{dist}$ mean/s.d. [ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td>I$\downarrow$-intoneme</td>
<td>-49/129</td>
<td>-48/146</td>
</tr>
<tr>
<td>N$\uparrow$-intoneme (pre-boundary)</td>
<td>115/156</td>
<td>90/197</td>
</tr>
<tr>
<td>N$\uparrow$-intoneme (other)</td>
<td>32/138</td>
<td>51/171</td>
</tr>
</tbody>
</table>

**Accent Command Timing and Duration.** Table 5.7 displays the relative timing for different intonemes. As can be expected, I$\downarrow$-intonemes are associated with early accent command timing, as by definition, the accent command offset is aligned with the accented syllable. The contrary is the case for N$\uparrow$-intonemes which start off later in the accented syllable. The standard deviations suggest a stronger connection between the syllable onset and the accent command onset, than between the accent command offset and the syllable offset. N$\uparrow$-intonemes preceding a prosodic boundary are associated with a considerably late accent command onset than others.

Table 5.8: Correlation between predictor variables and relative accent command timing.

<table>
<thead>
<tr>
<th>predictor variable p</th>
<th>$\rho(T1_{dist},p)$</th>
<th>$\rho(T1_{rel},p)$</th>
<th>$\rho(T2_{dist},p)$</th>
<th>$\rho(T2_{rel},p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>intoneme type</td>
<td>.51</td>
<td>.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>duration of syllable onset</td>
<td>.19</td>
<td>.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>predictor variable p</td>
<td>$\rho(T2_{dist},p)$</td>
<td>$\rho(T2_{rel},p)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>duration syllable coda</td>
<td>- .31</td>
<td>-.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>intoneme type</td>
<td>.38</td>
<td>.35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Table 5.7 the most important predictor factors with respect to accent command timing are displayed. For $T1$, these are the type of intoneme and the duration of the syllable onset. The importance of the latter parameter indicates that $T1$ is actually aligned with the onset of the nuclear vowel as shown in Section 4.2. In the case of $T2$, the intoneme type and the duration of the syllable coda are the most important factors. The longer the coda, the earlier the accent command offset with respect to the syllable offset. The lower correlations for the parameters expressing accent command timing relative to the syllable duration ($T1_{rel}$ and $T2_{rel}$) suggests that the accent command timing is more appropriately expressed with respect to syllable onset and offset time. The mean accent command duration very closely matches the mean syllable duration (mean/s.d.): 242 ms/127 ms vs. 240 ms/82 ms. This suggests that the assumption that accent commands align with the accented syllable, and not, for instance, with the timing of the superordinate phrase, is viable.

The problem of accurately predicting $Aa$, i.e. word prominence is addressed in more detail in the following section.
5.4.3.7 Predicting Prominence

As indicated, prominence in terms of accent command amplitude $Aa$ assigned to constituents in an utterance can only be predicted very coarsely from input information such as the part-of-speech and the type and position of an accent. Although generally speaking words can be roughly classified as content or function words, with the latter being accented only in contrastive contexts, assigning prominence to content words proves to be a difficult task. As shown in Table 5.6 verbs, especially in sentence-final position are less often accented than nouns. The average accent command amplitude assigned to accented verbs, however, does not significantly differ from that of nouns. This indicates that for predicting $Aa$ of a content word other factors need to be taken into account, such as the linguistic (syntactic and semantic) environment and pragmatic requirements. For this reason, in addition to the statistical evaluation of the entire corpus, a phrase-wise analysis was performed on half of the corpus [MJ01a]. In the following, a small number of accentuation patterns will be discussed that could be identified stably in the corpus. Instances of these patterns, however, are rather infrequent ($N < 100$) and therefore statistically weak compared with the number of 3022 accented syllables in the corpus.

**Enumerations.** Examining instances of lists containing three items, typically names, “A, B and C...”) yields higher prominence for the first and the third item than for the second, as in the examples shown in the following table ($Aa$ given in brackets):

<table>
<thead>
<tr>
<th>Laupheim, (.39)</th>
<th>Peisenberg (.22)</th>
<th>und Speyer (.35)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bośnien, (.29)</td>
<td>Kroatien (.23)</td>
<td>und Serben (.29)</td>
</tr>
<tr>
<td>Münch, (.50)</td>
<td>Perschau (.33)</td>
<td>und Schreiber (.46)</td>
</tr>
<tr>
<td>Deutschland (.36)</td>
<td>Frankreich (.26)</td>
<td>und die Niederlande (.32)</td>
</tr>
</tbody>
</table>

**Sequence of Function and Name.** The news stories very often refer to persons of public interest who are introduced with their function and name. In these constructions, the function is generally less prominent than the name. The following lists renders a few examples ($Aa$ given in brackets):

<table>
<thead>
<tr>
<th>Bundesaußenminister (.09) Kinkel (.33)</th>
</tr>
</thead>
<tbody>
<tr>
<td>der frühere (.24) Regierungschef (.00) Münch (.68)</td>
</tr>
<tr>
<td>Sachsen-Anhalts (.12) Regierungschef (.00) Höppner (.44)</td>
</tr>
<tr>
<td>der bosnische (.13) Außenminister (.00) Cecelje (.50)</td>
</tr>
</tbody>
</table>

**Given and New.** In this context ‘Given and New’ simply refers to whether or not a word (typically a name) has already been mentioned in the current news story. Consistent decrease of prominence can mostly be observed between first and second mention, especially when they occur in consecutive sentences. If the distance is larger, as for instance in a third mention, the word prominence is likely to increase again. The following lists renders a few examples ($Aa$ for first and second mention given in brackets):

<table>
<thead>
<tr>
<th>Carter (.32, .11)</th>
<th>Horstmann (.52, .20)</th>
<th>Scharping (.36, .22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burns (.59, .13)</td>
<td>Castro (.66, .24)</td>
<td>Masowiecki (.53, .36)</td>
</tr>
<tr>
<td>Rau (.82, .29)</td>
<td>Scharping (.35, .16)</td>
<td>Däubler-Gmelin (.37, .13)</td>
</tr>
</tbody>
</table>
The same pattern only applies to repeated mention of the same word. If a person is introduced by his name and later referred to by his function, a similar decrease of prominence is usually not observed.

The small selection of recurring accentuation patterns presented in the current section suggests possible limitations of a statistical approach to predicting prominences. Since instances of these patterns occur infrequently in the database, there might be too few for influencing the behavior of the neural network. Furthermore, this kind of infrequent, yet stable phenomena could more easily be taken care of by formulating prominence rules (for lists, for patterns of function-name, for second mentions etc.).

5.4.4 Pauses

Analysis shows that the duration of inter-phrase pauses is correlated with the accent command amplitude $A_p$ observed in these locations ($\rho = .57$). 83.0% of sentence boundaries ($BI = 4$) are connected with a pause, with a duration of 716 ms/336 ms (mean/standard deviation). Of the intra-sentence boundaries ($BI = 3$) only 52.6% are accompanied by a pause of 327 ms/132 ms.

5.4.5 Fujisaki Model Parameter vs. ToBI Labels

5.4.5.1 Introduction

In order to evaluate the correlation between Fujisaki model commands and ToBI labels, the former were examined with respect to their temporal position relative to accented and pre-boundary syllables (accent commands), and prosodic boundaries (phrase commands) [MF00]. Commands which had not been automatically associated with ToBI labels were looked up in the database and checked for their properties. In the case of accent commands, for instance, commands which could not be motivated by an accented syllable or boundary tone were deleted.

5.4.5.2 The Stuttgart G-ToBI System

We give a short introduction to the G-ToBI systems, as it differs from other formulation such as in [RB94]. The main tone types in the Stuttgart systems are L*H (a rise after a low accent syllable) and H*4L (a fall after a high accent syllable). Figure 5.23 to 5.25 show examples of these accents. The distribution of accent types over the whole database is displayed in Figure 5.27.

Break Indices ($BI$s) are distinguished according to the following properties:

<table>
<thead>
<tr>
<th>$BI$</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>word boundaries in clitic groups</td>
</tr>
<tr>
<td>1</td>
<td>phrase-medial word boundaries</td>
</tr>
<tr>
<td>2</td>
<td>disjuncture (pause) without clear tonal cue</td>
</tr>
<tr>
<td>3</td>
<td>intermediate phrase boundary</td>
</tr>
<tr>
<td>4</td>
<td>intonation phrase boundary</td>
</tr>
</tbody>
</table>

Table 5.9 lists the correspondences between MFGI and G-ToBI.

5.4.5.3 Results

Figure 5.26 displays an example of analysis, showing from top to bottom: the speech waveform, the extracted and model-generated $F_0$ contours, the ToBI tier, the text of the utterance, and the underlying phrase and accent commands.
5.4. STATISTICAL ANALYSIS OF THE PROSODIC DATABASE

Figure 5.23: Example of L*H accent.

Figure 5.24: Example of H*L accent.
Accent Assignment. The corpus contains a total number of 13151 syllables. Of the 2,498 syllables labeled asaccented 96.1% were found to be aligned with accent commands, as well as 78% of the 859 syllables assigned boundary tone labels. Accents immediately preceding a boundary with a mean accent command amplitude $A_a$ of .38 are found to be significantly stronger than non-boundary accents with a mean $A_a$ of .26.

‘Non-downstepped’ accents (98.0% of all accent labels) exhibit a mean accent command amplitude of .28 against .21 for accents labeled as down-stepped. Furthermore, accents marked as uncertain (‘?’, 1.9% of all accent labels) exhibit significantly lower $A_a$ than those labeled with certainty (.21 against .28). This indicates that it is the assessment of weaker accents that usually poses problems to the labeler.

The main standard accent types ‘H*L’,‘L*H’ which account for 81% of the accent labels can be reliably identified by the alignment of the accent command with respect to the accented syllable, expressed as $T_{1\, dist} = (T1 - t_{on})$; and $T_{2\, dist} = (T2 - t_{off})$ where $t_{on}$ and $t_{off}$ denote the accented syllable’s onset and offset time, respectively. For type ‘H*L’, mean $T_{1\, dist}$ and $T_{2\, dist}$ are -60 ms and -37 ms, and for type ‘L*H’ 132 ms and 168 ms, respectively. These results prove to be highly significant ($p < .01$). Figure 5.28 displays the accent command timing with respect to the accented syllable duration and the different accent types.

A considerable number of syllables ($N = 444$) exhibiting accent commands had not been assigned any accent labels by the human labeler. Figure 5.26 shows two such instances. In the utterance “Die Friedensgespräche für das ehemalige Jugoslawien im amerikanischen Dayton...” — “The peace talks for the former Yugoslavia in the American town of Dayton...”, accent commands were assigned to the words ‘ehemaligen’ and ‘amerikanischen’, but not tone labels. This indicates that labels are mainly missing where accents are relatively weak.

Phrase Boundaries. About 54.8% of $BI$ 3- and 96.2% $BI$ 4-labeled-boundaries are aligned with the onset of a phrase command, with a mean phrase command magnitude $A_p$ of .67 and 1.32, respectively.

Figure 5.25: Example of L*H concatenating with and H*L accent forming a hat pattern.
Table 5.9: Correspondences between MFGI and G-ToBI.

<table>
<thead>
<tr>
<th>MFGI</th>
<th>G-ToBI</th>
<th>intonational property</th>
</tr>
</thead>
<tbody>
<tr>
<td>phrase commands: Ap and T0,</td>
<td>BIs 0 to 4</td>
<td>phrase boundaries</td>
</tr>
<tr>
<td>accent commands: Aa, T1, T2</td>
<td>accent labels</td>
<td></td>
</tr>
<tr>
<td>N↑, I↓-intonemes</td>
<td>mostly L^H/</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H*L</td>
<td></td>
</tr>
<tr>
<td>accent command</td>
<td>boundary tone H%</td>
<td>boundary tone high</td>
</tr>
<tr>
<td>assigned to unaccented phrase-final</td>
<td>boundary tone L%</td>
<td>boundary tone low</td>
</tr>
<tr>
<td>syllable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>not labeled, but by default after I↓-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>intonemes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.26: An example of analysis from the database. The figure displays from top to bottom: (1) the speech waveform, (2) the extracted (+ signs) and estimated (solid line) F0 contours, ToBI labels and text of utterance, the underlying phrase commands (impulses) and accent commands (steps). The first phrase command with an Aa of 2.55 starts at $t = -0.165$ s (outside the panel). In the utterance “Die Friedensgespräche für das ehemalige Jugoslawien im amerikanischen Dayton...” — “The peace talks for the former Yugoslavia in the American town of Dayton...” the third accent command marks a minor accent on ‘ehemaligen’ — ‘former’ which was not assigned a ToBI-label.
Figure 5.27: Frequency of accent types in the database in %.

Figure 5.28: Timing of accent commands with respect to the accented syllable depending on the accent type.
Prosodic cues observed at boundaries include declination line resets — as triggered by phrase commands —, pauses, boundary tones and pre-boundary lengthening, the latter sometimes being the only cue at BI 3 prosodic boundaries. It must be stated, however, that the assignment of BIs by the labeler was sometimes inconsistent as boundaries with quite different prosodic cues and syntactic depths were assigned the same BI. If one compares the two boundaries assigned a BI of 4 in Figure 5.26, for instance, the second one precedes a long pause while the first one is merely marked with a boundary tone.

5.4.5.4 Conclusions
The comparison of Fujisaki parameters with the Stuttgart G-ToBI system indicates that both systems are compatible. The alignment of accent commands with the accented syllable corresponds to the definition of accent types. ‘Down-stepped’ accents exhibit smaller accent command amplitude Aa as could be expected. In terms of boundary depth, phrase commands and BI 4 boundaries mostly coincide, whereas BI 3 boundaries do not as often. Analysis of duration contours in these cases suggest that BI 3 boundary are often only marked by durational cues. It should be noted that the derivation of ToBI labels from Fujisaki parameters appears to be possible, but the loss of information incurred in terms of precise timing and amplitude is obvious. If one wishes to extract ToBI labels automatically, going through Fujisaki parameters for a good first approximation seems to be a viable approach.

5.5 Properties of the IGM
The current version of the IGM, which is based on the syllable as its basic rhythmic unit [MJ01a], outputs the prosodic parameters (1) syllable duration and (2) $F_0$ in terms of Fujisaki control parameters. A scaling factor for syllable intensity is calculated in parallel.

Figure 5.29 displays the output parameters of the IGM. For each syllable, the duration and, in the case of accented syllables and syllables bearing boundary tones, the parameters of the accent command assigned to the syllable, are calculated. Along with the amplitude Aa, the onset time $T_1$ and offset time $T_2$ of the accent command are output, the latter two relative to the onset and offset time of the syllable, respectively. If a syllable is the first in a prosodic phrase, the onset
Table 5.10: Output parameters and most important predictor variables. \( t_{on} \) and \( t_{off} \) denote onset and offset time of the current syllable, respectively.

<table>
<thead>
<tr>
<th>Output Parameter out of Model</th>
<th>Predictor Variable ( in ) of Model</th>
<th>( \rho(\text{out, in}) )</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>syllable duration</td>
<td>sum of duration means of phone classes in syllable</td>
<td>.64</td>
<td>13,151</td>
</tr>
<tr>
<td></td>
<td>boundary depth (right), 0=clinic, 1=word, 2=phrase, 3=sentence, 4=paragraph</td>
<td>.46</td>
<td>13,151</td>
</tr>
<tr>
<td></td>
<td>strength (0=unstressed, 1=stressed, 2=accented)</td>
<td>.35</td>
<td>13,151</td>
</tr>
<tr>
<td></td>
<td>nucleus schwa/non-schwa</td>
<td>-.19</td>
<td>13,151</td>
</tr>
<tr>
<td>( Aa )</td>
<td>type of intoneme (tone switch class)</td>
<td>.26</td>
<td>3,022</td>
</tr>
<tr>
<td></td>
<td>part-of-speech</td>
<td>.13</td>
<td>3,022</td>
</tr>
<tr>
<td></td>
<td>phrase index in sentence</td>
<td>-.12</td>
<td>3,022</td>
</tr>
<tr>
<td>( T1_{dist} = T1 - t_{on} )</td>
<td>type of intoneme</td>
<td>.51</td>
<td>3,022</td>
</tr>
<tr>
<td></td>
<td>number of phones in syllable onset</td>
<td>.15</td>
<td>3,022</td>
</tr>
<tr>
<td>( T2_{dist} = T2 - t_{off} )</td>
<td>type of intoneme</td>
<td>.38</td>
<td>3,022</td>
</tr>
<tr>
<td></td>
<td>number of phones in syllable rhyme</td>
<td>-.20</td>
<td>3,022</td>
</tr>
<tr>
<td>( A_p )</td>
<td>boundary depth (left)</td>
<td>.70</td>
<td>1,047</td>
</tr>
<tr>
<td></td>
<td>index of phrase in sentence</td>
<td>-.51</td>
<td>1,047</td>
</tr>
<tr>
<td></td>
<td>duration of preceding phrase</td>
<td>.32</td>
<td>1,047</td>
</tr>
<tr>
<td></td>
<td>( A_p ) of preceding phrase command</td>
<td>-.18</td>
<td>1,047</td>
</tr>
<tr>
<td></td>
<td>duration of current phrase</td>
<td>.11</td>
<td>1,047</td>
</tr>
<tr>
<td>( T0_{dist} = t_{on} - T0 )</td>
<td>distance from preceding phrase command</td>
<td>.26</td>
<td>1,047</td>
</tr>
<tr>
<td>intensity (mean frame power rms in syllable)</td>
<td>index of phrase in sentence</td>
<td>-.21</td>
<td>13,151</td>
</tr>
<tr>
<td></td>
<td>coda voiced</td>
<td>.14</td>
<td>13,151</td>
</tr>
<tr>
<td></td>
<td>index of syllable in phrase</td>
<td>-.12</td>
<td>13,151</td>
</tr>
<tr>
<td>pause</td>
<td>boundary depth (left)</td>
<td>.62</td>
<td>1,047</td>
</tr>
<tr>
<td></td>
<td>index of phrase in syllable</td>
<td>-.38</td>
<td>1,047</td>
</tr>
</tbody>
</table>

The time \( T0 \) of the phrase command assigned to the phrase is calculated with respect to the onset time of the syllable, as well as the phrase command magnitude \( A_p \). The speaker-dependent base frequency \( F_b \) and time constants \( \alpha \) and \( \beta \) are treated as constants. They are set to 50.2 Hz, .95/s and 20.3/s, respectively.

Phone duration is calculated from the superordinate syllable’s duration taking into account the phone properties found in the database. In the present implementation of the model, the syllable duration is repartitioned to the individual segments by adding up the segment means found in the database and linear scaling. Consonant segments are separately treated depending on whether they pertain to the syllable onset or the coda.

In order to capture potential interactions between intonation and rhythm, the prosodic parameters are predicted from a set of linguistic and phonetic input features using a single, feed-forward neural network (FFNN), since calculating syllable durations first and relating \( F_0 \) to these in a second step would still result in a sequential model. The design, optimization and training of the FFNN was performed by Oliver Jokisch, Dresden University of Technology.\(^1\)

Most other regression models, such as GLMs, CARTs, and sums-of-products models predict a single output factor from a set of input features, whereas FFNNs can predict various output parameters in parallel and have been shown capable of predicting prosodic parameters directly from raw observed data, as well as in terms of control parameters for the Fujisaki model.

As a summary of the preceding section, Table 5.10 lists the output parameters of the final

---

\(^1\)With regard to the details of the neural network design and training procedure we refer to the forthcoming D.Eng. thesis of Jokisch [Jok02], though basic considerations are already discussed in [JMO00].
5.5. PROPERTIES OF THE IGM

Figure 5.30: Example of analysis from the database. From top to bottom: speech waveform, extracted and model-generated $F_0$ contours, duration contour (syllabic $z$-score), ToBI tier, text of utterance, underlying phrase and accent commands. In the utterance “In der bosnischen Moslem-Enklave Bilacz…” — “In the Bosnian Muslim-envelope of Bilacz…” the accent command before the phrase boundary at ‘Bilacz’ exhibits a considerably high amplitude $A_a$.

IGM and their most important predictor factors along with the correlation coefficients as yielded by statistical analysis. It needs to be noted, that correlations for $A_a$, $T_{1_{dil}}$ and $T_{2_{dil}}$ were calculated only for accented syllables and syllables bearing boundary tones ($N = 3,022$), and correlations for $A_p$, $T_{0_{dil}}$ and pause (the duration of a pause preceding a prosodic phrase) for syllables which are the first in a prosodic phrase ($N = 1,047$). The parameter strength indicates whether a syllables is unstressed (0), stressed, but unaccented (1), or stressed and accented (2), i.e. bearing a tone switch. Prosodic boundaries are classified as intra-word / inter-word clitic (depth=0), inter-word (1), inter-phrase (2), inter-sentence (3, at full stops) and inter-paragraph (4, start of news story) boundaries.

From this overview it becomes clear, that the model incorporates information from lower level units (i.e. coda, rhyme, phones) as well as higher levels (word, phrase, sentence, paragraph) in the syllabic parameters. Relationships as rendered by the table are generally in line with the results of earlier works [Mix98, page 133 ff]. Comparison shows, that, especially in the case of $A_a$ and intensity, single input variables have relatively little predictive power, whereas for others, such as syllable duration and $A_p$ single parameters explain more than 40 % of the variance. As expected, a good predictor of syllable duration is the sum of mean durations of phone classes (in the database) pertaining to the syllable, with identical consonant phonemes being treated as different phone classes depending on their position in either coda or rhyme. In the case of $A_a$, the parameter reflecting the relative prominence given to an accented syllable, strong differences were found depending on whether or not an accent precedes an intra-sentence phrase boundary (mean of $A_a$ .34 against .25, compare example in Figure 5.30), whereas the part-of-speech of the superordinate word has relatively little influence. The apparently weak contributions of these parameters indicate, that additional information, such as the focal condition (narrow vs. wide focus) associated with an accent, is missing in the database, as well as a more detailed description of the syntactic environment.
20 + 4 (context) features

8 prosodic parameters
(integrated model)

Figure 5.31: Overview of FFNN-based IGM predicting eight parameters from a set of 24 input features.

### 5.6 Training and Testing the IGM

Based on the results of analysis discussed in the preceding section, 20 syllabic features were selected as syllable-based input vector (see Table 5.11 for reference), and augmented by four context parameters: (for accented syllables) the part-of-speech of the preceding accented word and the amplitude $Aa$ assigned to the preceding accent command, and, in the case of phrase-initial syllables, the properties of the preceding phrase commands ($T_0_{\text{dist}}, A_p$). The output vector consists of five Fujisaki model parameters with relative timing controlling the $F_0$ contour ($T_{1\text{dist}}, T_{2\text{dist}}, A_a, T_0_{\text{dist}}, A_p$), the current syllable duration and the duration of a potential pre-syllabic pause (syllable duration, pause), and also a signal intensity parameter. The training and prediction tasks are solved by a fully-connected feed-forward neural network (FFNN) of four layers ($24 \times 18 \times 12 \times 8$ neurons); Figure 5.31. Depending on the parameter ranges the input and output parameters are linearly scaled. Both, log and tan-hyperbolic transfer functions are used. The NN is trained using a teaching input (Stuttgart corpus) and standard error backpropagation, minimizing the root mean square error (RMSE) between teaching input and net output. The Stuttgart corpus was subdivided into a training set (10,000 syllables) and an independent test set (3,151 syllables). Observing the RMSE in the test set an over-adaptation to the training data was avoided even at a total of 500 – 1,500 training cycles. Although the network has a fairly simple structure, it is apparently suited to predict the eight output parameters of the IGM concurrently. The accuracy of prediction for individual parameters obviously depends on the predictive power of the 24 selected input parameters and the quality of the prosodic labels in the database from which they are computed (see Section 4). Table 5.12 shows the RMSE, means and standard deviations of trained and predicted output parameters.
Table 5.11: Complete list of input parameters of the IGM. The right column refers to the abbreviations used in Figure 5.31.

<table>
<thead>
<tr>
<th>Syllable Level Parameters</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>sum of mean durations of phones in syllable</td>
<td>OR_DUR</td>
</tr>
<tr>
<td>sum of mean durations of phones in onset</td>
<td>O_DUR</td>
</tr>
<tr>
<td>sum of mean durations of phones in rhyme</td>
<td>R_DUR</td>
</tr>
<tr>
<td>nuclear vowel schwa/non-schwa</td>
<td>SCHWA</td>
</tr>
<tr>
<td>number phones in onset</td>
<td>N_ON</td>
</tr>
<tr>
<td>Word Level Parameters</td>
<td>Abbreviation</td>
</tr>
<tr>
<td>index of syllable in word</td>
<td>IJN_WRD</td>
</tr>
<tr>
<td>duration class of part-of-speech of word</td>
<td>POS_DUR</td>
</tr>
<tr>
<td>number of syllables in word</td>
<td>SYLS_WRD</td>
</tr>
<tr>
<td>lexical word accent (0/1)</td>
<td>ACC</td>
</tr>
<tr>
<td>Parameters on the Phrase Level and above</td>
<td>Abbreviation</td>
</tr>
<tr>
<td>syllables in preceding phrase</td>
<td>SYL_PREC</td>
</tr>
<tr>
<td>boundary tone (0/1, before phrase boundaries)</td>
<td>BND</td>
</tr>
<tr>
<td>break index to the left (0–4)</td>
<td>BI_SYN_L</td>
</tr>
<tr>
<td>break index to the right (0–4)</td>
<td>BI_SYN_R</td>
</tr>
<tr>
<td>index of phrase in sentence</td>
<td>I_PHR</td>
</tr>
<tr>
<td>index of sentence in paragraph</td>
<td>L_SENT</td>
</tr>
<tr>
<td>start of phrase (0/1)</td>
<td>PHR_STA</td>
</tr>
<tr>
<td>start of paragraph (0/1)</td>
<td>PARA_STA</td>
</tr>
<tr>
<td>start of sentence (0/1)</td>
<td>SENT_STA</td>
</tr>
<tr>
<td>type of intoneme (three classes)</td>
<td>INTONENE</td>
</tr>
<tr>
<td>syllable strength (0–2)</td>
<td>STRENGTH</td>
</tr>
</tbody>
</table>

Table 5.12: RMSE between trained and predicted output parameters and their respective means and standard deviations.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>overall RMSE</th>
<th>mean(trained)</th>
<th>s.d. (trained)</th>
<th>mean (pred.)</th>
<th>s.d. (pred.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>syllab.dur.</td>
<td>.016 s</td>
<td>.189 s</td>
<td>.078 s</td>
<td>.191 s</td>
<td>.064 s</td>
</tr>
<tr>
<td>T1_dist</td>
<td>.058 s</td>
<td>.037 s</td>
<td>.152 s</td>
<td>.051 s</td>
<td>.097 s</td>
</tr>
<tr>
<td>T2_dist</td>
<td>.068 s</td>
<td>.039 s</td>
<td>.181 s</td>
<td>.052 s</td>
<td>.118 s</td>
</tr>
<tr>
<td>Aa</td>
<td>.06</td>
<td>.29</td>
<td>.17</td>
<td>.27</td>
<td>.06</td>
</tr>
<tr>
<td>T0_dist</td>
<td>.138 s</td>
<td>.435 s</td>
<td>.544 s</td>
<td>.432 s</td>
<td>.196 s</td>
</tr>
<tr>
<td>Ap</td>
<td>.14</td>
<td>1.10</td>
<td>.62</td>
<td>1.08</td>
<td>.57</td>
</tr>
<tr>
<td>pause dur.</td>
<td>.069 s</td>
<td>.290 s</td>
<td>.348 s</td>
<td>.290</td>
<td>.255 s</td>
</tr>
<tr>
<td>intensity</td>
<td>690</td>
<td>1697</td>
<td>775</td>
<td>1677</td>
<td>360</td>
</tr>
</tbody>
</table>
Table 5.13: Comparison of prediction results from the FFNN and single parameter linear regression models. The better correlation values are set in bold type.

<table>
<thead>
<tr>
<th>unit</th>
<th>$\rho(\text{meas.}, \text{pred.})$</th>
<th>$\rho(\text{meas.}, \text{pred.})$</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>syll. dur.</td>
<td>.81</td>
<td>.81</td>
<td>13,151</td>
</tr>
<tr>
<td>$T_1\text{dist}$</td>
<td>.61</td>
<td>.58</td>
<td>3,022</td>
</tr>
<tr>
<td>$T_2\text{dist}$</td>
<td>.63</td>
<td>.59</td>
<td>3,022</td>
</tr>
<tr>
<td>$A_a$</td>
<td>.40</td>
<td>.41</td>
<td>3,022</td>
</tr>
<tr>
<td>$T_0\text{dist}$</td>
<td>.53</td>
<td>.35</td>
<td>1,047</td>
</tr>
<tr>
<td>pause</td>
<td>.73</td>
<td>.76</td>
<td>1,047</td>
</tr>
<tr>
<td>intensity</td>
<td>.46</td>
<td>.37</td>
<td>13,151</td>
</tr>
</tbody>
</table>

It becomes clear that the predicted parameters exhibit reduced standard deviations, indicating the averaging behavior of the neural network. After re-scaling all parameters, and relating the relative timing parameters of the Fujisaki model to the timing of the underlying syllable chain, the model commands can be fed into the Fujisaki model, producing a time-aligned $F_0$ contour. The combination of a data-driven approach, such as a neural network, with a rule-based prosodic model can be considered as a hybrid architecture (HYDRA, see also [JH+98] which addresses the rapid adaptation to prosodic model parameters using a well-defined rule-based core).

Table 5.13 lists testing results expressed by the correlation between measured and predicted output parameters. It can be seen that correlations for syllable durations are considerably higher than for the Fujisaki control parameters, especially $A_a$. These results confirm the observations in the preceding section concerning possibly missing information in the set of predictor variables.

In order to conduct a baseline test as to whether an integrated prediction of prosodic parameters bears advantages over single-parameter methods, linear regression models for each output parameter were calculated from the complete set of input features, with correlations given in the second column from the right of Table 5.13.

Comparison between FFNN and regression-model based results indicates improvements in five parameters, especially timing parameters and intensity, and slight deterioration in the remaining three. This suggests certain synergy effects due to the integrated approach, but obviously requires further investigation.

5.7 Discussion and Conclusions

In the case of intensity and accent command amplitude $A_a$, obviously additional information needs to be taken into account, for the former supposedly the internal structure of the syllable, and for the latter focal condition and syntactic environment of an accented syllable. Considering the properties of the neighboring syllables as additional input parameters might as well improve the accuracy of the predictor.

Syllable Duration Figure 5.32 shows a histogram of the prediction error. About 67% of predicted syllabic durations are accurate to within 40 ms, with an RMSE of 46 ms and a mean of $-11$ ms.
5.7. DISCUSSION AND CONCLUSIONS

Although the correlation coefficient of .81 for syllable durations appears relatively low compared with the results of Zellner-Keller and Campbell, for instance, who report values between .85 and even .93, the following factors could explain the difference:

- The size of the corpus used in the current study of 13,151 syllables is considerably larger than the 1,204 and 3,959 used by Zellner-Keller and Campbell, respectively. More variation can be expected in a larger database.
- The speaking style (phonologically balanced sentences, short story in the case of Z.K. and C.) might well influence the durational variation.
- The Stuttgart database was segmented automatically and contains a relatively large number of transcription errors (segment identity, syllabification), whereas the other corpora were labeled manually.
- The internal structure of the syllable is only taken into account with respect to the durational mean values of phones involved. A more sophisticated modeling as proposed in [Cam93] might improve the overall performance of the model without changing the general concept.

Especially the latter factor might explain why the model yields better predictions on unknown sentences than on sentences from the database proper. This observation was made when
CHAPTER 5. AN INTEGRATED APPROACH TO MODELING PROSODY

predicting syllable durations for the perception experiment reported in Section 6.6 where correlations of over .86 were reached. Whereas the neural network successfully abstracts from errors in the database, the errors are reintroduced when comparing predicted and observed durations within the Stuttgart corpus. The largest single cause for predicted syllable durations being slightly shorter on the average, might be the erroneous transcription of the grapheme /e/ which in many cases is rendered as [e:] instead of the considerably shorter schwa [ə].

Fujisaki Model Parameters The correlations between predicted and observed Fujisaki parameters are difficult to interpret in terms of the quality of the resulting $F_0$ contour, since all parameters jointly influence $F_0$. For this reason, a perceptual evaluation as performed in the next chapter is required. It should be noted that the correlation between predicted and measured (raw) $F_0$ contours — depending on the individual utterance in the evaluation — was in the range between .58 and .65. This is a value that compares to the results reported by Paul Taylor for his Rise-Fall-Continuation model and the TILT model [Tay00, page 1707] of .62 - .65.

The relatively low correlations for $Aa$, however, suggest that the input information is not sufficient for predicting prominence. As we saw in Section 5.4.3.7 there is a number of possible additional factors which, however, occur only in relatively rare instances. In order to incorporate these factors in the statistical approach presented here they would need to be clustered or introduced to the model by some kind of post-hoc rule-system. If we consider the principles underlying accentuation in general, then whether to place an accent in a word and on which syllable is the most important decision, whereas the relative strength of the accent might be of lesser importance. Hence the linguistic pre-processing of the text is a very critical stage in the synthesis process, as the prosodic model cannot recover from errors committed and even highlights them.

Syllable Intensity Although there was no effort to optimize the model with respect to the syllable intensity, it is worthwhile comparing the results from the IGM with a work by Bagshaw [Bag98] who reports an RMSE of 6.8 dB for his model jointly predicting segment duration and intensity. The respective value for the IGM on syllables is 3.45 dB.

The following chapter will be dedicated to the perceptual evaluation of the IGM.
Chapter 6

Perceptual Evaluation of the Integrated Model

Abstract
This chapter discusses a series of perception experiments evaluating the prosodic quality of synthetic speech produced with the IGM. The comparison is performed using manipulated resynthesized original utterances and diphone synthesis examples. An important issue in this context are the relative contributions of syllable duration and $F_0$ to perceived naturalness.
6.1 Introduction

In the preceding section we already examined the accuracy with which the integrated model predicts its output parameters. Figures, such as the cross-correlation between measured and predicted parameters, \( \rho(nat, mod) \), and the rms error, however, only present relatively unspecific measures for assessing the quality of a prosodic model. Ultimately any model needs to be judged by human listeners. Perceptual experiments can also help us to determine the ‘perceptual quality distance’ between two approaches with different \( \rho(nat, mod) \), for instance, in order to develop machine-based methods for evaluating prosodic quality.

This chapter discusses a series of perception experiments evaluating the prosodic quality of synthetic speech produced with the IGM. The experiments were designed with the following questions in mind:

1. What are the relative contributions of syllable duration and intonation to perceived naturalness, and how does naturalness deteriorate when both parameters are jointly manipulated?

2. How close does the integrated model come to natural speech with respect to the prosodic quality?

3. How does the IGM perform on sentences not pertaining to the news corpus?

4. How does the IGM compare with the original rule-based approach (RBM), both in terms of durational and intonational quality?

5. How much does the segmental quality influence the quality judgement?

6. Can speech envelope manipulations according to predicted intensity values improve quality?

6.2 Experiment Design

Based on the experience from the experiments documented in Chapter 3 we decided to design an evaluation of the integrated prosody model consisting of the following parts:

**Experiment 1:** A grading experiment using resynthesized stimuli of 12 isolated sentences from the news corpus created with the IGM placed in a matrix of reference stimuli [MJ01c].

**Experiment 2:** A grading experiment using resynthesized stimuli of 12 isolated sentences, six from the news corpus and six from the corpus in Section 3.3.2.

**Experiment 3:** A grading experiment using the sentences from Experiment 2 with additional versions produced by the original rule-based model (RBM) evaluated in Chapter 3. Stimuli comprise resynthesis examples from original utterances as well as diphones synthesis (MBROLA) examples [MJ01b].

**Experiment 4:** A/B comparison of diphone synthesis stimuli only. One part of stimulus pairs is created for comparing IGM and RBM, the second part contains samples for examining the effect of intensity manipulations using IGM-predicted syllable intensity as a scaling factor.

Table 6.1 gives an overview of the experiments in terms of method, stimuli, and problems addressed in the experiments.
6.3 Selection of Test Sentences for the Evaluation

Twelve sentences from the news corpus were selected for performing the perceptual evaluation. These sentences were chosen by the following criteria: They should not exceed a certain length (duration < 10 s) and complexity (one or two syntactic phrases), in order not to overstrain the subjects' short term memory.

For easier processing they were selected from a subset of five different news stories. These sentences will henceforth be referred to as the ‘D set’ (Table 6.2).

In order to evaluate the performance of the IGM on sentences not pertaining to the news corpus (Experiments 2, 3, and 4), six sentences were taken from the corpus of isolated sentences used in the perceptual evaluation in section 3.3.2 which are listed in Table 6.3.

Henceforth these sentences which were produced by a single male speaker, will be referred to as the ‘J set’.

6.4 Creating a Matrix of Reference Stimuli

As we saw in Section 3.4, comparing speech synthesis to recorded speech as a reference only yields a very coarse assessment of the prosodic quality, as the segmental deterioration incurred by concatenating diphones is immense. For this reason we decided to maintain the segmental quality on a level as high as possible and only manipulate prosodic features. Hence, departing from the natural utterance and by controlled degrading in the durational and P0 dimensions it is possible to create resynthesized stimuli with a high segmental quality and a specific prosodic quality, as close to or distant from natural speech as desired. Figure 6.1 is meant to illustrate this approach. The vertical arrows indicate possible quality judgements.

Therefore, as a reference matrix and in order to examine the relative contributions of durational and intonational quality to the overall judgement, an array of stimuli was created by
Table 6.2: List of sentences in the D set.

| D01 | Dort setzte heute morgen eine Flüchtlingswelle ein. |
| D02 | Die französischen Experten arbeiteten mit größter Umsicht. |
| D03 | Deshalb sei auch die geäußerte Kritik unfair. |
| D04 | Der Verband südostasiatischer Staaten, ASEAN, hat heute auf seiner Jahrestagung im Sultanat Brunei Vietnam aufgenommen. |
| D05 | ASEAN will vor allem die wirtschaftliche Zusammenarbeit fördern. |
| D06 | Der kubanische Staats- und Parteichef Castro will an der Macht bleiben. |
| D07 | Heute müsse man die beeindruckenden Erfolge Chinas, Vietnams sowie der Koreanischen Volksdemokratischen Republik als vorbildlich zur Kenntnis nehmen. |
| D08 | Der Imperialismus sei heimtückisch wie immer. |
| D09 | Gegen sie werde Anzeige wegen schweren Landfriedensbruchs erstattet. |
| D10 | Die Sicherheitskräfte in der Innenstadt wurden weiter verstärkt. |
| D11 | In der vergangenen Nacht war es erneut zu Ausschreitungen gekommen. |
| D12 | Die serbische Bevölkerung ziehe in Treks nach Südwesten, hieß es. |

Table 6.3: List of sentences in the J set.

| J01 | Bereitwillig gab er Auskunft. |
| J02 | Aller Anfang ist schwer. |
| J03 | Die Begründung ist stichhaltig. |
| J06 | Das Gespräch zeigte die Gegensätze und die gemeinsamen Züge unserer Auffassungen. |
| J07 | Wenn wir die Maschine anschließen, beginnt der Motor zu surren. |
| J08 | Es regnete soviel, dass der Fluss über die Ufer trat. |
controlled degrading of the prosodic features of the natural utterances. The degree of degradation was defined by the cross-correlation between natural and modified/predicted parameters. In the case of syllable durations, degradation was achieved by incremental compressing or stretching, yielding overall target cross-correlations $\rho_{dur}$ of .90, .80, .70, and .60. These target correlations were chosen on the basis of preliminary tests which had also indicated that correlations for the $F_0$ contours needed to be considerably lower in order to yield comparable effects. All stimuli were created by applying the PSOLA resynthesis functionality of the software Praat (v.3.8.68 [BW96]) and replacing DurationTiers and PitchTiers.

The degrading of the $F_0$ contours was not performed on the individual $F_0$ values, but by modifying the Fujisaki control parameters estimated from the natural utterances. Since micro-prosody is absent in the smoothed $F_0$ contours produced by the Fujisaki model, the average cross-correlation $\rho_{F_0}$ between original and modeled contours is of .93, measured for frames of 10 ms. Departing from the original Fujisaki parameter configurations, by incremental reduction or increase of parameter values, $F_0$ contours were created with cross-correlation coefficients $\rho_{F_0}$ of .70, .50 and .30, while observing that $Aa$ and $Ap$ cannot assume negative values.

In order to equally distribute the error across all control parameters involved ($T_1_{dist}, T_2_{dist}, Aa, T0_{dist}, Ap$), the increment was adjusted individually for each parameter until the parameter-based cross-correlation between original and modified parameters was nearly equal, while observing the target value for the $F_0$ contour-based cross-correlation. The center panels of Figure 6.2 display examples of $F_0$ contours for $\rho_{F_0}$ of .70 and .50.

### 6.5 Experiment 1: The IGM in a Matrix of Reference Stimuli

#### 6.5.1 Introduction

The first experiment was designed for testing the integrated model within a relatively fine grid of reference stimuli. A main purpose was to determine the relative contributions of duration and $F_0$ prediction accuracy to perceived naturalness. Therefore the experiment also aimed at testing the viability of prosodically degrading natural speech samples and establish the relationship between correlation coefficients of prosodic parameters and the naturalness expressed by a mean
Figure 6.2: Four examples of $F_0$ contours. The top panel displays from top to bottom: the original speech waveform, the extracted (+) and model-generated $F_0$ contours (solid line), duration contour (syllabic z-score), ToBI tier, text of utterance, underlying phrase and accent commands. Utterance displayed: "In der bosnischen Muslem-Enklave Bihac gingen die Kämpfe zwischen den Regierungstruppen..." - "In the Bosnian Muslim-enclave of Bihac, fighting between the government troops...". The lower three panels display $F_0$ contours and underlying Fujisaki model commands for the following cases: (a) degraded Fujisaki parameters, $\rho_{E_0} = .70$; (b) degraded Fujisaki parameters, $\rho_{E_0} = .50$; (c) Fujisaki parameters predicted by the integrated model. It can be seen that the amplitudes of predicted accent commands (bottom panel) show less variation than the extracted ones (top panel).
6.5. EXPERIMENT 1: THE IGM IN A MATRIX OF REFERENCE STIMULI

opinion score.

6.5.2 Setting

Subjects taking part in the experiment were 21 students of Media Computer Sciences at Berlin University of Applied Sciences, 17 males and 4 female, in their second year. They were informed that the experiment dealt with the quality of synthetic speech, but not about the details of parameters manipulated. Figure 6.3 displays a map of the stimuli created for the experiment. It becomes clear that not all possible combinations of durational and $F_0$ modifications were tested. As the experiment was performed during a scheduled lecture, a maximum experiment duration of 90 minutes had to be observed.

![Figure 6.3: Array of stimuli used in Experiment 1. The dark-gray triangles denote versions using integrated model-based prediction, one with predicted durations and natural $F_0$, one with natural durations and predicted $F_0$, and a third one using both durations and $F_0$ from the prediction. The light-gray stars denote stimuli produced using extracted Fujisaki parameter configurations.](image)

Informal listening tests had shown little degradation for stimuli with extracted Fujisaki parameters ($\rho_{F_0} \approx .93$), as well as for those with durational correlations of $\rho_{dur} = .90$. Therefore in these cases only three combinations were created, otherwise five. For the stimuli generated using the integrated model, an average $\rho_{F_0}$ of .55 was calculated, and an average $\rho_{dur}$ of .81. As can be seen from Figure 6.3, for each sentence, 25 different versions were created, yielding a total number of 300 stimuli. In order to test the consistency of the quality judgement, stimuli pertaining to four of the sentences were included twice, bringing the total number of stimuli to 400. The subjects were provided with forms and requested to assess the quality of the stimuli with grades between 1 (very good) and 5 (very bad), according to the German grading system. Intermediate grades 2, 3 and 4 were explained as corresponding to judgements of ‘good’, ‘acceptable’ and ‘bad’.

As a brief introduction to the ‘quality spectrum’ of the stimuli, and also in order to familiarize the subjects with the voice characteristics of the speaker, an original recording of a news story was played back (supposed to receive a judgement of ‘1’). In the following a random choice of 10 stimuli with manipulated prosodic parameters of varying quality was presented. During the assessment phase of the experiment, stimuli were presented in randomized order and played back twice for every decision, while observing that consecutive stimuli pertained to different sentences. After presenting the first half of the stimuli, a five-minute break was taken.

The grading approach which obviously holds the risk of less accurate judgements compared with A/B comparisons as performed in [MMH99], for instance, was chosen because of the large
number of stimuli involved, since A/B comparison between 25 different stimuli would imply 25 \times 24/2 = 300 decisions times the number of sentences, and even A/B comparison between adjacent stimuli only would have implied a number of about 100 decisions per sentence. With a total number of 400 stimuli presented in 90 minutes, and considering the numerous repetitions of the same sentence, the cognitive load on the students was extremely high. Especially the stimuli with extremely low correlation values of $\rho_{dur} = .60$ and $\rho_{F_0} = .30$ which had been introduced to the set of utterances in order to mark the ‘very bad’ end of the quality spectrum, sometimes raised amused reactions because of their strongly distorted prosody.

### 6.5.3 Results

In order to yield a score increasing with perceived stimulus quality (MOS), grades assigned by the subjects were subtracted from a value of 6, producing a scale increasing with perceived quality from 1 to 5. The MOS averaged over all subjects and sentences is displayed in Table 6.4. The general tendency that the MOS monotonically decreases with deteriorating prosody becomes obvious, with the steepest decay found between a $\rho_{dur}$ of .70 and .60. The natural speech stimuli were not unanimously rated ‘very good’ and only reach a MOS of 4.37, with an average of 7.5 of the 16 ‘original’ stimuli in the set of utterances being assigned ‘grade 1’.

Furthermore it can be seen that the scores for the stimuli produced with the integrated model exceed those assigned to adjacent ‘degraded stimuli’. The MOS for the integrated model (predicted durations, natural $F_0$) of 4.18 compares to that of $\rho_{dur} = 1.00 / \rho_{F_0} = .90$ (4.13). In the case of integrated model-based $F_0$ and natural durations, the score of 3.83 compares to $\rho_{dur} = 1.00 / \rho_{F_0} = .70$ (3.82). Joint predictions of duration and $F_0$ (3.40) is perceptually closest to the ‘degraded stimuli’ of $\rho_{dur} = .80 / \rho_{F_0} = .90$ (MOS = 3.33).

Factor analysis, excluding stimuli produced with the integrated model, showed a correlation between the MOS and $\rho_{dur}$ of .79, and a correlation between MOS and $\rho_{F_0}$ of .29 ($p < .01$ for both factors), indicating that duration is the predominant factor for the quality judgement. The influence of one parameter on the MOS, however, depends on the quality of the other, as can be seen from Table 6.5. The correlation between $\rho_{F_0}$ and MOS drops with decreasing $\rho_{dur}$.

The identity of the sentence was identified as a secondary factor ($p < .05$). No significant correlation, however, was found between the MOS and the length of a sentence in terms of the number of syllables it contains.

Although all subjects actually made use of all available grades from 1 to 5, the individual averages vary between 2.14 and 3.85 (mean = 2.87, slightly less than the scale mean of 3.0). The mean inter-subject correlation was found to be of .63, indicating considerable individual variations. The correlation between ratings at first and second time of presentation amounts to

<table>
<thead>
<tr>
<th>$\rho_{F_0}$</th>
<th>1.00</th>
<th>.90</th>
<th>.70</th>
<th>.50</th>
<th>.30</th>
<th>IGM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{dur}$</td>
<td>4.37</td>
<td>4.13</td>
<td>3.82</td>
<td>3.66</td>
<td>3.50</td>
<td>3.83</td>
</tr>
<tr>
<td>1.00</td>
<td>3.94</td>
<td>-</td>
<td>3.54</td>
<td>-</td>
<td>3.29</td>
<td>-</td>
</tr>
<tr>
<td>.90</td>
<td>3.57</td>
<td>3.33</td>
<td>3.17</td>
<td>2.96</td>
<td>2.79</td>
<td>-</td>
</tr>
<tr>
<td>.70</td>
<td>3.07</td>
<td>-</td>
<td>2.73</td>
<td>2.57</td>
<td>2.53</td>
<td>-</td>
</tr>
<tr>
<td>.60</td>
<td>2.35</td>
<td>2.10</td>
<td>2.04</td>
<td>1.93</td>
<td>1.88</td>
<td>-</td>
</tr>
<tr>
<td>IGM</td>
<td>4.18</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.40</td>
</tr>
</tbody>
</table>
6.6. EXPERIMENT 2: SENTENCES NOT PERTAINING TO THE NEWS CORPUS

Table 6.5: Correlation between $p_{F_0}$ and MOS, for constant $p_{dur}$ (left), and correlation between $p_{dur}$ and MOS for constant $p_{F_0}$ (right).

<table>
<thead>
<tr>
<th>$p_{dur}$</th>
<th>$\rho(p_{F_0}, \text{MOS})$</th>
<th>$p_{F_0}$</th>
<th>$\rho(p_{dur}, \text{MOS})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.100</td>
<td>.62</td>
<td>.100</td>
<td>.88</td>
</tr>
<tr>
<td>.90</td>
<td>.61</td>
<td>.90</td>
<td>.84</td>
</tr>
<tr>
<td>.80</td>
<td>.55</td>
<td>.70</td>
<td>.82</td>
</tr>
<tr>
<td>.70</td>
<td>.51</td>
<td>.50</td>
<td>.81</td>
</tr>
<tr>
<td>.60</td>
<td>.41</td>
<td>.30</td>
<td>.77</td>
</tr>
</tbody>
</table>

.92 suggesting a high intra-subject consistency.

6.5.4 Discussion

Results indicate that subjects are far more sensitive to errors in the prediction of syllable durations than to variations in the $F_0$ contour. If one of the two parameters deteriorates, however, the influence of the other on the MOS weakens. This means, the better the duration prediction becomes, for instance, the more a poor $F_0$ prediction is noticed. In this context it must be noted that, since the $F_0$ contour is computed using alignment information based on syllabic durations, manipulations in the durational domain cause time-warping in the $F_0$ contour and hence further degradation.

In terms of predicted syllable durations, the integrated model comes close to natural durations whereas the prediction of Fujisaki control parameters, especially $A_a$, obviously requires additional input parameters. This already became evident with the results of statistical analysis discussed in Section 5.6.

6.6 Experiment 2: Sentences not Pertaining to the News Corpus

6.6.1 Introduction

The second experiment of the perceptual evaluation focused on the performance of the integrated prosodic model on sentences which do not pertain to the news corpus. For this purpose, six sentences were chosen from the D set used in Experiment 1 and complemented by the six sentences of the J set.

6.6.2 Setting

The setting was similar to Experiment 1. Stimuli produced with IGM were presented in a reference matrix of degradation stimuli. As for the versions created with the IGM, variants with predicted durations and $F_0$ contour, as well as variants with either of the features copied from the natural utterances were produced. Degradation stimuli were created with a minimum of $p_{dur}$ of .70, and a minimum of $p_{F_0}$ of .50. These values were chosen as the stimuli with lower correlation coefficients used in Experiment 1 had sounded overly distorted. Figure 6.4 displays the total matrix of stimuli employed.

$$G_p(t) = \begin{cases} 
\alpha^2t \exp(-\alpha t), & \text{for } t \geq 0, \\
0, & \text{for } t < 0.
\end{cases}$$  \hspace{1cm} (6.1)
In order to produce the degraded stimuli, the original recordings from the J set were analyzed using the automatic approach for Fujisaki parameter extraction. For the speaker of the J set, an \( F_b \) of 75 Hz and an \( \alpha \) of 2/s was determined. Creating the IGM-based stimuli for the J set posed certain problems: As the \( \alpha \) of the news speaker was only .95/s, and \( F_b \) as low as 50.2 Hz, the phrase command amplitudes \( A_p \) output by the IGM were considerably higher than those required for an \( \alpha \) of 2/s. This relationship results from the formulation of the phrase control mechanism: It is proportional to the square of \( \alpha \), Equation 6.1. Hence, for a given maximum \( F_0 \) modification caused by the phrase component (i.e. the log \( F_0 \) reached on the top of the phrase component, see Figure 6.5), a higher \( \alpha \) requires a lower \( A_p \) and vice versa. Preserving an \( \alpha \) of 2/s and re-scaling the phrase command amplitudes \( A_p \) with a constant factor did not solve the problem, as consecutive commands (due to the steeper slope at a higher \( \alpha \)) were too low to compensate for the declination effect. However, when \( F_b \) and \( \alpha \) were both taken from the news speaker D, the resulting \( F_0 \) contours were too low for speaker J.

Eventually, a compromise was found by keeping an \( \alpha \) of 0.95 /s and scaling \( A_p \) to match an \( F_b \) of 75 Hz. The scaling factor was empirically determined by adjusting the mean \( F_0 \) of the model-generated contours to that of the extracted natural contours.

Syllable durations for the J set were calculated using the IGM by supplying the underlying linguistic and phonetic information, i.e. the 24 input parameters of the IGM. Calculation showed that the correlation between measured and predicted syllable durations on the J set was of .86.

Subjects taking part in the experiments were 20 members of staff and students of the Laboratory of Acoustics and Speech Communication, Dresden University of Technology, 16 males and four females. All of them were experienced in listening to synthetic speech. For logistic reasons, unlike Experiment 1, Experiment 2 was performed on a PC using headphones. After being questioned their names, subjects were presented the randomized stimuli. Each stimulus was played back twice, then the subjects had to specify a grade between 1 and 5 (very good..very bad). The choices were automatically logged to a protocol file. The sequence of the first 30 stimuli was presented twice, in order to familiarize the subjects with the quality spectrum of the stimuli.
6.6. EXPERIMENT 2: SENTENCES NOT PERTAINING TO THE NEWS CORPUS

6.6.3 Results

Table 6.6 lists the mean opinion scores averaged over all subjects and the total of twelve sentences for the different stimulus conditions. As in Experiment 1, the MOS was determined by subtracting the grades assigned by the subjects from 6, yielding a scale ascending with quality between 1 and 5.

As can be seen from the table, the MOS monotonously decreases with $\rho_{dur}$ and $\rho_{F_0}$, but the effect is much stronger in the direction of $\rho_{dur}$. This is also reflected by the correlation coefficients between MOS and $\rho_{dur}$ and $\rho_{F_0}$, which are of .78 and .36, respectively. In the case where syllable durations and $F_0$ contours are supplied by the integrated model, it is judged slightly better (MOS= 2.99) than reference stimuli with $\rho_{dur}$ of .80 and $\rho_{F_0}$ of .70. If syllable

Table 6.6: Overview of MOS results averaged over all 12 sentences and all subjects.

<table>
<thead>
<tr>
<th>$F_0 \rightarrow$</th>
<th>1.00</th>
<th>.90</th>
<th>.70</th>
<th>.50</th>
<th>IGM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{dur}$ ↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>4.66</td>
<td>4.50</td>
<td>3.97</td>
<td>3.70</td>
<td>3.58</td>
</tr>
<tr>
<td>.90</td>
<td>4.40</td>
<td>4.02</td>
<td>3.80</td>
<td>3.51</td>
<td>-</td>
</tr>
<tr>
<td>.80</td>
<td>3.70</td>
<td>3.36</td>
<td>2.94</td>
<td>2.89</td>
<td>-</td>
</tr>
<tr>
<td>.70</td>
<td>2.70</td>
<td>2.63</td>
<td>2.30</td>
<td>2.17</td>
<td>-</td>
</tr>
<tr>
<td>IGM</td>
<td>4.13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.99</td>
</tr>
</tbody>
</table>
Table 6.7: Overview of MOS results for J set (left) and D set (right) listed separately.

<table>
<thead>
<tr>
<th>( \rho_{F_0} \rightarrow \rho_{d_{VAR}} )</th>
<th>1.00</th>
<th>.90</th>
<th>.70</th>
<th>.50</th>
<th>IGM</th>
<th>( \rho_{F_0} \rightarrow \rho_{d_{VAR}} )</th>
<th>1.00</th>
<th>.90</th>
<th>.70</th>
<th>.50</th>
<th>IGM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>4.55</td>
<td>4.26</td>
<td>3.69</td>
<td>3.40</td>
<td>3.06</td>
<td>1.00</td>
<td>4.77</td>
<td>4.73</td>
<td>4.26</td>
<td>4.00</td>
<td>4.11</td>
</tr>
<tr>
<td>.90</td>
<td>4.33</td>
<td>3.80</td>
<td>3.55</td>
<td>3.18</td>
<td>-</td>
<td>.90</td>
<td>4.46</td>
<td>4.24</td>
<td>4.05</td>
<td>3.85</td>
<td>-</td>
</tr>
<tr>
<td>.80</td>
<td>3.61</td>
<td>3.23</td>
<td>2.85</td>
<td>2.71</td>
<td>-</td>
<td>.80</td>
<td>3.79</td>
<td>3.49</td>
<td>3.02</td>
<td>3.07</td>
<td>-</td>
</tr>
<tr>
<td>.70</td>
<td>2.56</td>
<td>2.47</td>
<td>2.27</td>
<td>2.06</td>
<td>-</td>
<td>.70</td>
<td>2.84</td>
<td>2.80</td>
<td>2.32</td>
<td>2.28</td>
<td>-</td>
</tr>
<tr>
<td>IGM</td>
<td>3.81</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.57</td>
<td>IGM</td>
<td>4.44</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.40</td>
</tr>
</tbody>
</table>

durations only are taken from the IGM (MOS = 4.13), it is rated closer to a \( \rho_{d_{VAR}} \) of .90. IGM-predicted \( F_0 \) contour values only (MOS = 3.58) are rated even lower than the corresponding reference stimuli for a \( \rho_{F_0} \) of .50. The subjects obviously judged \( F_0 \) contours produced by the IGM rather critically.

If we display the results for D set and J set separately (see Table 6.7), it becomes clear that the samples from the J set were generally assessed to be of poorer quality (\( \rho \text{(MOS, corpus)} = .20, p < .01 \)). This observation not only concerns stimuli produced with the IGM, but even the originals. The quality distance, however, is greatest when the \( F_0 \) contours are supplied by the IGM. This suggests that especially the contours for the J set were judged inadequate by the listeners.

Furthermore, the results for the D set correspond better to those obtained in Experiment 1 (compare with Table 6.4).

6.6.4 Discussion

The results yielded in the current experiment suggest that the IGM, at least as far as the prediction of \( F_0 \) is concerned, performs poorer on 'unknown' sentences than on sentences from the news corpus on which it was trained. This could partly be attributed to the compatibility problem discussed in Section 6.6.2, as the \( \alpha \) chosen is considerably smaller than that of the original speaker.

A further reason for the dispreference of the J set might be the fact that the speaker of the J set himself was a member of staff of the Dresden laboratory for many years, and produced speech corpora and an inventory for diphone synthesis. Hence his voice quality and register were extremely familiar to all of the subjects in Experiment 2. The assumption that this 'conditioning' influenced the listeners' judgement is supported by the results of Experiment 3 discussed in the following section.

6.7 Experiment 3: The IGM Compared with the Original RBM

6.7.1 Introduction

The third experiment aimed at comparing the original sequential rule-based prosodic model evaluated in Chapter 3 with the integrated approach.

Although the rule-based model (RBM) is discussed in detail in Sections 2.6.2.2 (for the duration part) and 2.7 (for the intonation part), we will give a short summary of its properties and state the main differences between the IGM and RBM.

In the RBM phone duration is calculated using a Klatt-style formula in which departing
from a phone-specific inherent duration by subsequent application of rules factors such as the proximity of phrase boundaries or the presence/absence of lexical stress are taken into account for weighting the inherent duration. If a phone is part of a function word, for instance, it is compressed, but never below a phone-specific minimum value. The degree to which a phone can be compressed is defined by an elasticity factor. For each phone, the phonetic properties of adjacent phones are also taken into account. Based on the predicted phone durations, following a set of heuristic rules, Fujisaki model commands are aligned with the segmental string. Accent commands are set up according to the assignment of tone switches to accented or pre-boundary syllables (boundary tones). The fine alignment of accent command is realized with the onset of the nuclear vowel of an accented syllables as a reference point. In the case of rising accents, the onset of an accent command is aligned with this reference, and in the case of falling accents, the offset of an accent command. The respective offset and onset times are chosen considering the offset/onset of following/preceding nuclear vowels and a minimum criterion for accent command duration. In the case of accents close to utterance-medial boundaries, the boundary itself determines the accent command offset time. The amplitude $A_0$ is set with respect to the type of intoneme. Phrase commands are aligned with phrase boundaries, typically by setting up the phrase command at a distance of $1/\alpha$ before the segmental onset of the phrase, for the phrase component to reach its maximum at the segmental onset. The phrase command magnitude $A_p$ is set depending on the boundary depth and the number of syllables in the preceding phrase.

Hence the IGM and the RBM differ in the following respects:

- The rule-based model calculates phone durations based on a Klatt-style formula.

- According to the phone durations determined, the $F_0$ contour is aligned with nuclear vowels, based on a set of heuristic rules ([Mix98], p. 238 ff.) for determining the underlying Fujisaki parameters. These rules were established by analysis of a small corpus and optimized with respect to the TTS system DreSS.

- The IGM conjunctly calculates syllable durations and syllable-aligned Fujisaki parameters using an FFNN trained on data from a larger, prosodically labeled speech corpus.
• In the IGM, phone durations are calculated with respect to the superordinate syllable duration.

It should be noted, however, that the models do not substantially differ in the input information they process. Linguistic pre-processing and phonetic transcription employed is identical for both models.

6.7.2 Setting

As we saw in Section 6.5.2, for the stimuli generated using the integrated model, an average $\rho_{F_0}$ of .55 was calculated, and an average $\rho_{dur}$ of .82. For the rule-based model, an average $\rho_{F_0}$ of .54 was calculated, and an average $\rho_{dur}$ of .71.

The selection of sentences used in Experiment 3 was identical to that of Experiment 2: Stimuli were created for six sentences from the D set and the six sentences of the J set.

In addition, a selection of diphone synthesis stimuli was produced using the freely available MBROLA diphone synthesizer ([D'96], male voice de2) using the prosodic information for the resynthesis conditions. As a reference, originals and a selection of ‘degraded stimuli’ from the first two experiments were also included in the evaluation. The complete selection of stimuli is displayed in Figure 6.6.

Subjects taking part in the experiment were 21 students of Telecommunication Engineering at Berlin University of Applied Sciences in their fourth year. They were informed that the experiment dealt with the quality of synthetic speech, but they were not given details of parameters manipulated. The experiment took about 60 minutes to perform, with a ten-minute break after the first half of the stimuli. For each sentence 16 versions were created yielding 192 different stimuli. In order to test the consistency of the quality judgement, the stimuli from the prosodic models to be compared were included twice, bringing the total number of stimuli to 264. The subjects were provided with forms and requested to assess the quality of the stimuli with grades between 1 (very good) and 5 (very bad), according to the German grading system. Intermediate grades 2, 3 and 4 were explained as corresponding to judgements of ‘good’, ‘acceptable’ and ‘bad’. The sequence of the first sixteen samples in the first turn was presented twice, in order to familiarize the subjects with the ‘quality spectrum’ of the stimuli. During the assessment phase of the experiment, stimuli were presented in randomized order and played back twice for every decision, while observing that consecutive stimuli pertained to different sentences.

6.7.3 Results

In order to yield a score increasing with perceived stimulus quality (MOS), grades assigned by the subjects were subtracted from a value of 6, producing a scale from 1 to 5. The MOS averaged over all sentences and subjects depending on the approach is displayed in Figure 6.7. As can be seen, the natural speech stimuli were not unanimously rated ‘very good’ and only reach a MOS of 4.45, with an average of 59 % of the ‘original’ stimuli in the set of utterances being assigned ‘grade 1’. The second best rating of 3.93 was assigned to stimuli with durations produced with the integrated model and natural $F_0$. The corresponding value for the rule-based model is 3.14. A similar distance (3.17 vs. 2.48) is found for stimuli where durations as well as $F_0$ contours were predicted by the models. Comparison with stimuli using natural durations and model-based $F_0$ (3.66 vs. 3.67), however, indicates that the quality difference between the models must be attributed to the better duration prediction of the integrated approach, whereas the $F_0$ contours from the rule-based model are judged even slightly better. Differences between the prosodic models are very much leveled in the case of diphone based stimuli (1.73 vs. 1.63) which were unanimously placed at the lower end of the quality spectrum. The performance of
the integrated model on the J set was found to be slightly poorer that on the D set, as can be seen from the corpus-wise results for all versions of stimuli displayed in Table 6.8, the difference, however, is not as vast as in Experiment. In contrast, the RBM performs better on sentences from the J set than from the D set.

Although all subjects actually made use of all available grades from 1 to 5, the individual averages vary between 2.44 and 3.56 (mean = 2.88, slightly less than the scale means of 3.0). The mean inter-subject correlation was found to be of .61, indicating considerable individual variations.

6.7.4 Discussion

The results of Experiment 3 indicate that the integrated model outperforms the rule-based model mainly by the quality of segment durations. This outcome was likely because of the considerably higher $\beta_{dur}$ of the IGM.

In terms of the quality of the $F_0$ contours, however, the IGM is not judged significantly better than the RBM. The fact that the RBM performed better on the sentences of the J set could partly be explained by the fact that the RBM had been optimized on the diphone inventory produced by the speaker of the J set.

The results of statistical analysis presented in Section 5.4 suggested that features currently derived from plain text are not sufficient for accurately predicting prominences, i.e. the accent command amplitude $A_a$ assigned to accented items in an utterance. For this reason, the integrated model does not gain any improvement over the rule-based model in terms of the $F_0$ contours predicted. As discussed in Section 5.4.3.7, prominence is influenced by a number of
addition of infrequent, but stable factors. The current prediction which is simply based on position, accent type or part-of-speech captures only a fraction of the necessary information. It also needs to be questioned whether statistical approaches, currently being ‘state-of-the-art’, are capable of modeling infrequent events, and how they could be complemented by a set of rules.

The experiment outcome also makes clear that introducing resynthesis and diphone synthesis stimuli in the same experiment automatically places the latter on the lower end of the quality spectrum where segmental quality outrules prosodic quality. For this reason, Experiment 4 was designed which exclusively employs diphone synthesis stimuli.

6.8 Experiment 4: IGM vs. RBM (Diphone Synthesis), Intensity.

6.8.1 Introduction

The fourth experiment consisted of two kinds of comparison which were performed in parallel for the sake of simplicity:

1. The comparison between the IGM and the RBM on diphone synthesis examples.

2. The comparison between stimuli manipulated or unmanipulated with respect to their intensity.

The first comparison was designed as a follow-up to Experiment 3. As we saw from the results in Section 6.7.3, the diphone synthesis-based stimuli were assessed as being of poor quality and differences between the prosodic models which had been obvious in the resynthesis cases, were leveled. By comparing diphone-based examples only, we hoped to establish a comparable ranking between IGM and RBM for diphone synthesis.

As a by-product of the prosodic model, syllable-based intensity values are output. Although no effort was made to optimize the performance of the IGM on the intensity prediction, we were interested in whether or not using the output values for scaling the diphone synthesis examples had any effect on the perceived naturalness. Employing intensity scaling in resynthesis had not been attempted, as the speech signal already possessed a ‘natural’ envelope. In diphone synthesis, however, diphone units are recorded maintaining a nearly constant maximum recording level throughout all units. When units are concatenated, the volume of the units - at least in the MBROLA approach - is not manipulated.
6.8. EXPERIMENT 4: IGM VS. RBM (DIPHONE SYNTHESIS), INTENSITY.

6.8.2 Setting

As we were concerned that the grading approach applied during the first three experiments might fail because of the poor segmental quality of the diphone examples, we decided to perform an A/B comparison type of experiment.

For the comparison between IGM and RBM, we employed the same diphone synthesis stimuli that had been used in Experiment 3 comprising 12 sentences with the following combinations:

- Natural durations and IGM-based $F_0$
- IGM-based durations and $F_0$
- Natural durations and RBM-based $F_0$
- RBM-based durations and $F_0$

Hence, 24 pairs with one sample from the two approaches each were created.

In order to manipulate the speech envelope, syllable-based intensity values from the IGM for all utterances were converted into PRAAT IntensityTier format by aligning the syllable intensity point with the center of the respective syllable. As the IntensityTier format requires values in dB, these values were calculated by applying $20 \times \log_{10}$ to the original intensity values from the IGM. In order to create a second version with more strongly manipulated intensities, $30 \times \log_{10}$ was applied to the IGM intensity values. Although this yields higher nominal intensity values, PRAAT automatically rescales the speech signal when it is multiplied with an IntensityTier, and therefore the signal dynamics are increased without incurring distortions by clipping whatsoever.

Altogether the following versions were created:

- Natural durations and IGM-based $F_0$
  - no intensity manipulation (IGM-I0)
  - $20 \times \log_{10}$ manipulation (IGM-I20)
  - $30 \times \log_{10}$ manipulation (IGM-I30)

- Natural durations and extracted Fujisaki parameter-based $F_0$
  - no intensity manipulation (0.9-I0)
  - $20 \times \log_{10}$ manipulation (0.9-I20)
  - $30 \times \log_{10}$ manipulation (0.9-I30)

Pairs were created by combining one unmanipulated and one manipulated version each from identical duration/$F_0$ conditions, yielding a total of 48 pairs.

The stimuli were randomized assuring that consecutive examples pertained to different sentences, and played back over loudspeakers. The subjects were first presented sample A, then sample B and had to decide which one they perceived as sounding more natural, or if both sounded about as equally natural/unnatural.

Subjects taking part in the experiment were 8 students of Telecommunication Engineering and 10 of Media Computer Science at Berlin University of Applied Sciences. They were informed that the experiment dealt with the quality of synthetic speech but were not given details of parameters manipulated. The experiment took about 25 minutes to perform.
6.8.3 Results

**IGM vs. RBM**  The results are in line with the outcome of Experiment 3: The IGM outperforms the RBM only when $F_0$ and the syllable durations are produced by the prosodic models. Under this condition, the IGM is preferred over the RBM in 64% of cases, in 4% it is judged as equally natural as the RBM. If the syllable durations are copied from the original speech signal, the IGM is preferred only in an average of 46% of cases and dispreferred in 48%. This ratio, however, was not the same throughout subjects, as some of them preferred the IGM in up to 75% of cases, and others the RBM. This suggests that the models were 'recognized' by some of the subjects by their characteristic intonation patterns.

In this context, it is worthwhile noting that the 'equal' judgement was made only in a very small number of cases. This behavior of the subjects indicates that they indeed perceived noticeable differences between the two prosodic models.

**Intensity Manipulation**  In 34.5% of cases the unmanipulated examples were judged as sounding more natural than the I20 ($20 \times \log_{10}$ manipulation) stimuli, in 44.5% of cases as equally natural. In comparison with the I30 ($30 \times \log_{10}$ manipulation) stimuli, these figures are of 52.4% and 32.4% respectively. Hence, on the average, the intensity scaling did not yield any improvement. Especially some of the I30 stimuli sounded, as some of the subjects remarked, as if somebody was turning the volume of the loudspeakers high or low.

Only in the case of sentences D08, J02 and J07, the I20 stimuli were assessed slightly better than the unmanipulated ones, with percentages (equal/better/worse) of 52/39/9, 61/30/9 and 71/25/4, respectively.

Figure 6.8 displays speech waveforms from J02, “Aller Anfang ist schwer”, from top to bottom: The original, the original $F_0$ contour, the diphone synthesis with unmanipulated envelope (0.9–10), and the cases 0.9–I20 and 0.9–I30. When comparing the lower three panels the scaling effect becomes clear. Especially the last syllable of the utterance ‘schwer’ ($t = 0.92 - 1.40$ s) which has a comparably high amplitude in the unmanipulated case is attenuated and therefore closer to the original. In the case of the fourth syllable ‘-fang’ ($t = 0.43 - 0.71$ s) however, the scaling has an adverse effect: The syllable is attenuated in contrast to its relatively high amplitude in the original. It should be noted that very often in the case of rising accents (see $F_0$ contour of the original), post-accent syllables exhibit higher amplitudes than the accented syllable proper.

6.8.4 Discussion

The experiment confirms that the IGM produces more accurate syllable durations than the RBM, but in terms of $F_0$ contour prediction, the model does not perform significantly better. Intensity manipulations using the intensity parameter from the IGM do not noticeably improve the naturalness of the diphone synthesis. Although the intensity decrease towards the end of an utterance is captured by the intensity contour, in other cases the manipulation produces undesired amplitude modulations. This outcome suggests that more accurate prediction of syllable intensities requires additional input parameters to the integrated model.

6.9 Discussion and Conclusions

An interesting question with respect to the method of using reference stimuli was whether or not they were judged stably throughout the first three experiments, regardless of other stimuli being added or removed. Six sentences of the D set had been tested in all three experiments...
Figure 6.8: Examples of envelope scaling for sentence J02, “Aller Anfang ist schwer”. From top to bottom: The original, the original $F_0$ contour, the diphone synthesis with original intensity and the scaled versions 0.9-I20 and 0.9-I30.
Table 6.9: Overview of MOS and Standard Deviation for the part of the D set common to the first three experiments.

<table>
<thead>
<tr>
<th>$\rho_{\text{dur}}$</th>
<th>$\rho_{F_0}$</th>
<th>Exp. 1 MOS</th>
<th>Exp. 1 s.d.</th>
<th>Exp. 2 MOS</th>
<th>Exp. 2 s.d.</th>
<th>Exp. 3 MOS</th>
<th>Exp. 3 s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGM</td>
<td>IGM</td>
<td>3.20</td>
<td>.32</td>
<td>2.99</td>
<td>.64</td>
<td>3.17</td>
<td>.34</td>
</tr>
<tr>
<td>IGM</td>
<td>1.00</td>
<td>4.35</td>
<td>.38</td>
<td>4.13</td>
<td>.56</td>
<td>3.93</td>
<td>.33</td>
</tr>
<tr>
<td>.70</td>
<td>.50</td>
<td>2.55</td>
<td>.30</td>
<td>2.17</td>
<td>.43</td>
<td>2.21</td>
<td>.32</td>
</tr>
<tr>
<td>.70</td>
<td>1.0</td>
<td>2.93</td>
<td>.41</td>
<td>2.70</td>
<td>.27</td>
<td>2.80</td>
<td>.45</td>
</tr>
<tr>
<td>.80</td>
<td>.90</td>
<td>3.32</td>
<td>.26</td>
<td>3.36</td>
<td>.36</td>
<td>3.33</td>
<td>.36</td>
</tr>
<tr>
<td>.90</td>
<td>.70</td>
<td>3.63</td>
<td>.32</td>
<td>3.80</td>
<td>.51</td>
<td>3.51</td>
<td>.24</td>
</tr>
<tr>
<td>1.00</td>
<td>IGM</td>
<td>3.92</td>
<td>.31</td>
<td>3.58</td>
<td>.77</td>
<td>3.66</td>
<td>.26</td>
</tr>
<tr>
<td>1.00</td>
<td>.50</td>
<td>3.89</td>
<td>.25</td>
<td>3.70</td>
<td>.58</td>
<td>3.59</td>
<td>.42</td>
</tr>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>4.55</td>
<td>.20</td>
<td>4.66</td>
<td>.26</td>
<td>4.45</td>
<td>.24</td>
</tr>
</tbody>
</table>

under resynthesis conditions. For these sentences, the properties in terms of $\rho_{\text{dur}}$ and $\rho_{F_0}$ and the corresponding MOS values are listed in Table 6.9.

Statistical analysis yields no significant influence of the property of the experiment on the MOS, whereas the correlations between MOS and $\rho_{\text{dur}}$ and $\rho_{F_0}$ are .90 and .34, respectively. These results indicate that the method of using a reference matrix of degraded stimuli for assessing is viable and yields stable results.

Returning to the main questions at the beginning of this chapter we can state the following:

1. The accuracy of syllabic duration prediction is by far a more crucial factor than the accuracy of the $F_0$ contour. Subjects are more sensitive to rhythmical errors than inadequacies in the $F_0$ contour. Decorrelating syllabic durations to a value of .80, for instance, has a comparable effect on the MOS as decorrelating the $F_0$ contour to approximately .40 (compare Table 6.4). When the two parameters are manipulated jointly, the influence of one parameter decreases as the other deteriorates. In other words, the more improvement we yield in the prediction of syllabic durations, for instance, the more accurate our $F_0$ contour prediction must become.

2. In terms of duration prediction only, the IGM comes close to decorrelated stimuli with $\rho_{\text{dur}}$ of .90, for $F_0$ prediction only, the corresponding value for $\rho_{F_0}$ is .70. Joint prediction of both parameters places the IGM about 1 point lower than the original recording (on a 4 point scale).

3. As far as the performance of the IGM on sentences not pertaining to the news corpus is concerned, Experiments 2 and 3 yield somewhat controversial results. In Experiment 2, the ‘unknown’ sentences in the J set were rated considerably poorer when produced by the IGM (more than 1 point). Subjects taking part in Experiment 2 were very familiar with speaker J. In Experiment 3, the differences between the J set and the D set were not as vast (about .2 points less for the J set). As already discussed above, this outcome may partly be explained by the properties of the speakers who exhibited different $F_0$ and $\alpha$.

4. Experiments 3 and 4 showed that the IGM outperforms the RBM only in terms of its rhythmical quality whereas the $F_0$ contours produced by the models are assessed as being equally acceptable. Furthermore the RBM performs considerably better on the J set than
6.9. DISCUSSION AND CONCLUSIONS

on the D set. This outcome may be explained by the fact that the RBM had been optimized for an inventory produced by speaker J, whereas the IGM was trained on data to which the D set pertained.

5. Experiment 3 indicates that diphone based stimuli were rated about 3 points poorer than the originals and 2 points below the resynthesis stimuli. In this experiment, for the diphone samples, no significant difference was established between the prosodic models. This outcome indicates that ultimately segmental quality overrides prosodic quality. When diphone-based stimuli only are compared such as in Experiment 4, the ranking of the prosodic approaches on resynthesized samples is confirmed.

6. As for the current implementation of the IGM, no significant quality improvement was established for stimuli with manipulated envelopes. This outcome, however, does not prove that intensity manipulation in general could not improve perceived naturalness; it only shows that the IGM was not yet optimized for this purpose. At least in terms of modeling the intensity reduction observed towards the end of an utterance, the model behaves appropriately.

In conclusion it must be stated that there is much room for improvements in the current version of the IGM.

With respect to the accuracy of syllabic durations, a vital step forward will be the correction of transcription errors and segmental realignment of the prosodic database. Especially, the confusion between the [ɛ] and [æ] phones which is due to imperfect transcription rules for the grapheme /ɛ/ leads to noticeable shortenings of [ɛ] phones. Moreover, the canonical transcription generally applied in the database does not take into account vowel reductions and syllable mergers (compare, for instance, /-zio-/ becoming monosyllabic [tisjo:] and not the canonical bisyllabic [tio:]. This problem, for instance, led to segmental distortions in some of the resynthesized IGM-based stimuli.

As far as improvement of the $F_0$ prediction is concerned, the current input information does not appear sufficient for predicting word prominence. It must be questioned, however, whether or not additional linguistic parameters exist which can actually raise the performance of the statistical model.

The method of using degraded stimuli as a reference matrix has proven to be a suitable concept for testing prosodic models. However, the parameters $\rho_{\text{dur}}$ and $\rho_{F_0}$ of the degraded stimuli can only be seen as very coarse quality measures. Moreover, the relative contributions of degraded individual Fujisaki parameters to the perceived quality of the resulting $F_0$ contour is not clear. In the current approach, for simplicity, all five parameters employed, i.e. $Aa$, $T1_{\text{dist}}$, $T2_{\text{dist}}$, $Ap$, $T0_{\text{dist}}$, were decorrelated to the same extent by adding or subtracting appropriately scaled increments. By only degrading individual parameters, say, for instance, only $Ap$, however, an indefinite number of different sounding $F_0$ contours could be produced which all share the same $\rho_{F_0}$, but not necessarily yield the same MOS.
Chapter 7

Discussion and Conclusions

This work presented an integrated approach to modeling German prosody. It should be noted that the line of research presented here is somewhat off the mainstream of activities in this field. At present there are relatively few studies employing the Fujisaki model for parametrizing \( F_0 \) contours, especially in European languages, and the model - for one exception known to the author - is not used in any existing TTS systems for European languages.

One criticism frequently made is that the Fujisaki model lacks a linguistic basis and parameters cannot be straightforwardly estimated from natural \( F_0 \) contours. The former argument is somewhat off the target, as there was never a claim that the model presents a linguistic representation. The model permits to calculate an \( F_0 \) contour from a quasi-discrete command-wise representation which by Analysis by Synthesis can be inferred from the natural \( F_0 \) contour. We developed a robust method for extracting the parameters from the natural \( F_0 \) contour and have shown in this work that the commands and their timing and amplitudes can be related to the linguistic contents of an utterance.

Departing from a model of German intonation based on the Fujisaki formula, MFGI, that was the topic of the author’s D.Eng. thesis, a perceptual evaluation of the \( F_0 \) model for speech synthesis was performed.

Evidence was found that, although MFGI performs better than other approaches, the temporal grid from a suboptimal duration model limits the prosodic naturalness of the synthetic speech issued from the TTS system. This observation was made when natural segment durations were employed in the synthesis and yielded significantly higher naturalness ratings.

As a consequence, an integrated model of prosody was proposed and developed based on the consideration that speech synthesis lacks the coherence of the natural speech signal which is partly due to the sequential approach to modeling prosody. Of the currently available statistical models for predicting prosody only neural network based approaches are capable of predicting several parameters in parallel. This property was utilized for training a network jointly predicting syllable duration and intensity, and \( F_0 \) in terms of Fujisaki model command parameters.

For evaluating the integrated approach a novel method for controlled prosodic degrading of natural speech was developed. This method was used for creating a grid of reference stimuli degraded in the durational as well as in \( F_0 \) dimension. As was shown by the results, the mean opinion score is strongly correlated with the correlation of prosodic parameters, with syllabic duration playing the far more important role.

The evaluation of the model performance yields the following main results:

- The joint prediction of prosodic parameters indicates certain synergy effects in the case of certain output parameters of the model over a single-parameter approach, but these are not significant for all parameters. Enforced optimization effort into the single-parameter
approaches might well lead to different results.

- Stimuli created with the integrated model are rated better than degraded stimuli of comparable $\rho_{dur}$ and $\rho_{F_0}$, but still about one point below natural stimuli on a four-point scale.

- The integrated model outperforms the original rule-based model only in terms of syllable duration prediction, the $F_0$ contour are rated even slightly poorer, especially for sentence from outside the news corpus.

In terms of the original assumptions underlying this work it must therefore be stated that the integrated approach does not seem to possess significant advantages over single-parameter methods, except for the fact that it leads to a very compact prosodic model. A full-blown optimization of multi-parameter versus single-parameter approaches was beyond the scope of this thesis, but would have probably demonstrated only marginal differences.

The assumption that the accuracy of duration prediction is crucially important was confirmed both by the results of the perceptual evaluation of degraded stimuli and by the comparison between the original sequential rule-based model and the integrated approach.

From all the output parameters of the model $Aa$ showed the weakest correlation between predicted and observed parameters. Here lies supposedly the major problem of this approach (and probably all current approaches to $F_0$ prediction in TTS [vS97a]). As we have shown in Section 4.3, $Aa$ is deeply linked with perceived prominence, whose relative assignment to constituents in an utterance creates meaning. The semantically impoverished information we extract from text which is used for predicting the prosodic parameters in our models is not sufficient and can therefore not account for the variance in prominence we see in the natural $F_0$ contour. This dilemma, of course, is not a shortcoming of the $F_0$ parametrization which we use, but an inherent problem of unrestricted speech synthesis from text altogether. Current models can produce naturally sounding $F_0$ contours, but they still do not give the listener the impression that the computer knows what it is saying. In Section 5.4.3.7 we indicated that the detailed analysis of prominence using the Fujisaki parametrization permits us to identify regular patterns which, however, occur relatively seldom throughout a corpus and are most probably speaking style dependent. It remains a challenge to incorporate these observations in a speech synthesizer.

What are the implications? The author is convinced that prosody research will continue, and it is still in need of well-defined models. Yet current trends seem to point in the opposite direction. On the one side we see approaches promoted which follow the philosophy to touch the speech signal and therefore its prosody as little as possible (non-uniform unit data-driven speech synthesis). On the other side we are told that we need to add all the information which we cannot extract from the text by developing an elaborate mark-up system. Neither approach — in the mind of this author — seems worth investing research time and effort. Joining a future army of database labelers and corporate voice designers cannot be the ultimate challenge of speech research. Prosody is.
Bibliography


Appendix A

Glossary of Special Terms Used in the Thesis

**HFC.** The High Frequency Contour denotes the faster changing components of an interpolated and smoothed $F_0$ contour and is derived by high-pass filtering ($f_g = 0.5$ Hz).

**IGM.** Throughout this thesis, IGM denotes an integrated model for prosodic control within MFGI which jointly computes all prosodic parameters, that is, segment durations, the $F_0$ contour in terms of Fujisaki control parameters, pause durations and syllable intensity from a set of twenty linguistically and phonetically motivated input parameters. Since the modeling unit is the syllable, all Fujisaki control parameters are related to the syllable bounds. Phone durations are derived from the superordinate syllable duration.

**LFC.** The Low Frequency Contour denotes the slower changing components of an interpolated and smoothed $F_0$ contour and is derived by subtracting the output of a high-pass filter ($f_g = 0.5$ Hz), the HFC, from the original smooth $F_0$ contour.

**MFGI.** The Mixdorff-Fujisaki Model of German Intonation presents a special adaptation of the per-se language-independent Fujisaki model of the production process of $F_0$ to German. Based on earlier concepts of German intonation, namely the tone switch approach by Isačenko and the description of German sentence intonation by Stock, criteria are developed for applying the Fujisaki model to $F_0$ contours of German and for interpreting the model commands in terms of the underlying linguistic and para-linguistic information in an utterance. Criteria mainly concern the association of accent commands with basic intonational elements, so-called intonemes and boundary tones, and the correspondence between phrase commands and major intonation phrases. Intonemes, boundary tones and phrase boundaries can be derived by application of a set of intonation rules to a given text. Applied to TTS, MFGI consists of a two-stage process: (1) the generation of symbolic prosody from text, that is, the appropriate sequence of intonemes and boundary tones, and the potential locations of phrase boundaries, (2) the generation of the $F_0$ contour proper from the symbolic prosody. For processing stage (2), two possible ways have been examined, a rule-based sequential model for first calculating segment durations and then the $F_0$ contour (RBM), and an integrated statistical approach which jointly calculates segment durations and the $F_0$ contour (IGM).

**RBM** Throughout this thesis, RBM stands for the sequential rule-based model for deriving segment durations and the $F_0$ contour used in the DRESS TTS system evaluated in Chapter 3.
In this conventional approach, phone durations are first calculated using a Klatt-style recursive rule system. The $F_0$ contour is then generated also by rule, by calculating phrase and accent commands which are aligned with the segments.
Appendix B

Corpus of Sentences used in Perception Experiments

The following sentences were used in the perception experiments discussed in Chapter 3 and Sections 6.6 – 6.8 [Meh85].

1. Statement Intonation

1.1 Du sagst es. You say it.
1.2 Ihr ruft uns. You call us
1.3 Wir hören. We’re listening.
1.4 Heinz spielt. Heinz is playing.
1.5 Hören Sie. Listen!
1.6 Schwer ist das. This is difficult.
1.7 Ihr lest. You’re reading.
1.8 Es rauscht. There’s a hiss.
1.9 Prüft es genau! Examine it thoroughly!

1.10 Ihr beschreibt es. You’re describing it.
1.11 Du untersuchst es. You’re examining it.
1.12 Er war bei den Jüngsten. He was with the youngest.

1.13 Wir erzählen es. We’re telling it.
1.14 Wir freuen uns daran. We’re looking forward to it.
1.15 Ihr hörte ihn doch schon. But you heard him already.

1.16 Ihr lest das Buch. You’re reading the book.
1.17 Wir treffen uns morgen. We meet tomorrow.
1.18 Der Akzent ist wichtig. The accent is important.

1.19 Wir hören genau zu. We’re listening carefully.
1.20 Übung macht den Meister.  Practice makes perfect.
1.21 Bereitwillig gab er Auskunft. He readily supplied information.
1.22 Aller Anfang ist schwer. It’s always difficult at the beginning.
1.23 Die Begründung ist stichhaltig. The explanation is valid.
1.24 In dem Lehrbuch sind viele Hinweise enthalten. The textbook contains many pieces of advice.
1.25 Üben und immer wieder üben ist beim Erlernen jeder fremden Sprache notwendig. Practicing again and again is necessary for learning any foreign language.
1.26 Das Gespräch zeigte die Gegensätze und die gemeinsamen Züge unserer Auffassungen. The conversation made clear the controversies and common grounds of our views.
1.27 Wir haben eine dankbare Aufgabe. We have a rewarding task.
1.28 Ist diese Lampe aber schön. Isn’t this lamp beautiful.
1.29 Rauchen verboten! No smoking!
1.30 Laßt uns weitergehen! Let’s go on!
1.31 Bedenken Sie bitte seine Bescheidenheit! Please consider his humbleness!
1.33 Zangen, Feilen und Bohrer lagen auf dem Tisch. Pliers, files and drills were lying on the table.
1.34 Wir lachten, sangen und tanzten vor Freude. We laughed, sang and danced for joy.
1.35 Wir betrachten die Größe, Gestalt und Anordnung der Kristalle. We studied the size, shape and arrangement of the crystals.

2. Non-terminal Intonation

2.1 Du mußt alle kennen, wenn Du sie beurteilen willst. You need to know all of them if you want to judge them.
2.2 Wir fragen ihn gern um Rat, denn er weiß meistens eine Lösung. We like to ask his advice as he usually knows an answer.
2.3 Ich weiß wirklich nicht, ob dies die richtige Methode ist. I really don’t know if this is the right method.
2.4 Wenn wir die Maschine anschließen, beginnt der Motor zu summen. When we connect the machine, the motor starts humming.
2.5 Als wir die Maschine angeschlossen hatten, begann der Motor zu summen. When we had connected the machine, the motor started humming.
2.6 Sobald der Dirigent den Takstoch hob, trat tiefe Stille ein. As soon as the conductor raised his baton, deep silence entered.
2.7 Wir stimmen dem Vorschlag zu, weil Ihr uns überzeugt habt. We agree to the proposal, because you convinced us.
2.8 Er arbeitete unermüdlich, um die Höhe seiner Kunst zu erreichen. He worked untiringly to reach the height of his art.
2.9 Es regnete soviel, daß der Fluß über die Ufer trat. It rained so heavily that the river burst its banks.
3. Question Intonation

3.1 Hört Ihr es?       Do you hear it?
3.2 Verstehst Du mich? Do you understand me?
3.3 Ist es kalt draußen? Is it cold outside?
3.4 Haben Sie es gehört? Did you hear it?
3.5 Wo liegt das Buch?  Where lies the book?
3.6 Wo das Buch liegt?  Where the book lies?
3.7 Wo bildet sich Kieselsäure? Where does silicic acid develop?
3.8 Wo sich Kieselsäure bildet? Where silicic acid develops?
Appendix C

Acknowledgements

I am deeply indebted to Prof. emer. Hiroya Fujisaki (University of Tokyo) for his scientific guidance and support of this work, which is based on the results of the almost seven years of our cooperation in intonation research. Furthermore I would like to thank Prof. Rüdiger Hoffmann (TU Dresden) especially for the moral and logistic support of my research efforts and for providing a ‘scientific homebase’ for me in Dresden. I thank Prof. John Local (University of York) for an inspiring stay at his lab and his constructive criticism.

Special thanks go to Prof. Dieter Mehnert (Berlin), my co-worker on a two-years project, who supported me with his experience of several decades of activity in the field of prosody, and also provided invaluable advice concerning the layout of this thesis.

I would like to thank Prof. Wolfgang Hess (University of Bonn) for his continued support.

Credits also go to Oliver Jokisch (TU Dresden), Maria Wolters (formerly University of Bonn, now Rhetorical Systems, Edinburgh) and Christina Widera (University of Bonn) for their cooperation in joint projects, and to Stefan Werner (University of Joensuu/Finland) who provided a quiet place for wrapping up my thesis at his lab in the summer of 2001.

The larger part of this work was supported by the DFG (Deutsche Forschungsgemeinschaft), grants no. HO 1674 and MI 625.

Last, but not least I want to thank my wife Noi for her understanding and care.
Appendix D

Curriculum Vitae

Personal Details
Date and Place of Birth: 28 March 1964, Berlin
Marital Status: Married
Children: Ken and Tim

Career History

Education

1970-1976  Primary school in Berlin
1983      Abitur
January-March 1983 Practical training for electrical engineering students, AEG, Berlin
1983-1990  Technical University Berlin, studies of electrical engineering
1990      Graduated, Diplom-Ingenieur
            Specialized in software engineering, digital speech processing
            and radio-frequency engineering

1990-1993  Free University Berlin, studies of Japanese, Chinese and phonetics
January 1993  
Granted a two-year post-graduate scholarship by the Japanese Ministry of Education and the German Academic Exchange Service (DAAD)

April-September 1993  
Osaka University of Foreign Studies, Japanese language course

October 1993-March 1995  
Science University of Tokyo with Prof. Fujisaki, studies in German and Japanese intonation, analysis of $F_0$ contours using the Fujisaki model, rule-based generation of $F_0$ contours of German, perception experiments.

June 1996  
Granted a two-year DFG-funding for a research project on prosodic control in speech synthesis at TU Dresden

May 1998  
Doctor of Engineering (TU Dresden)

January-March 1999  
DFG-funded research project at the National Chiao Tung University, Hsinchu (Taiwan)

January 2000-March 2001  
DFG post-doc fellow

February/March 2002  
DAAD short-term lecturer and researcher at Chulalongkorn University, Bangkok (Thailand)

**Employment**

1987-1990  
During university holiday: Working student with AEG-Sendertechnik, Berlin. Computer simulation of shortwave antennas, software development

1987-1990  
Student assistant in the interdisciplinary research project “Real time visualization of speech for the hearing impaired”, Institute for Communication Science, Technical University of Berlin. Hard- and software development for speech processing purposes.

April-September 1990  
Development engineer with Telefunken-Sendertechnik (former AEG). Design of shortwave antennas and networks, software development.

October 1990-March 1993  
Research engineer (part time) in the successor project to “Real time visualization of speech...”. Development of a real-time pitch extractor/spectrum analyzer and other tools for speech processing.

November 1995-December 1999  
Research engineer with TELES AG, Berlin. Development of speech technology (speech recognition / synthesis) for telephony applications, head of speech group

April 1999-March 2001  
Lecturer of Digital Signal Processing at the Faculty of Electrical Engineering, TFH-Berlin University of Applied Sciences

Since April 2001  
Professor of Digital Audio and Video Processing at the Faculty of Computer Science, TFH-Berlin University of Applied Sciences
Further skills and interests

Languages: German (mother tongue), English (fluent spoken and written), French, Spanish, Russian, Japanese, Chinese and Thai (all fluent spoken)