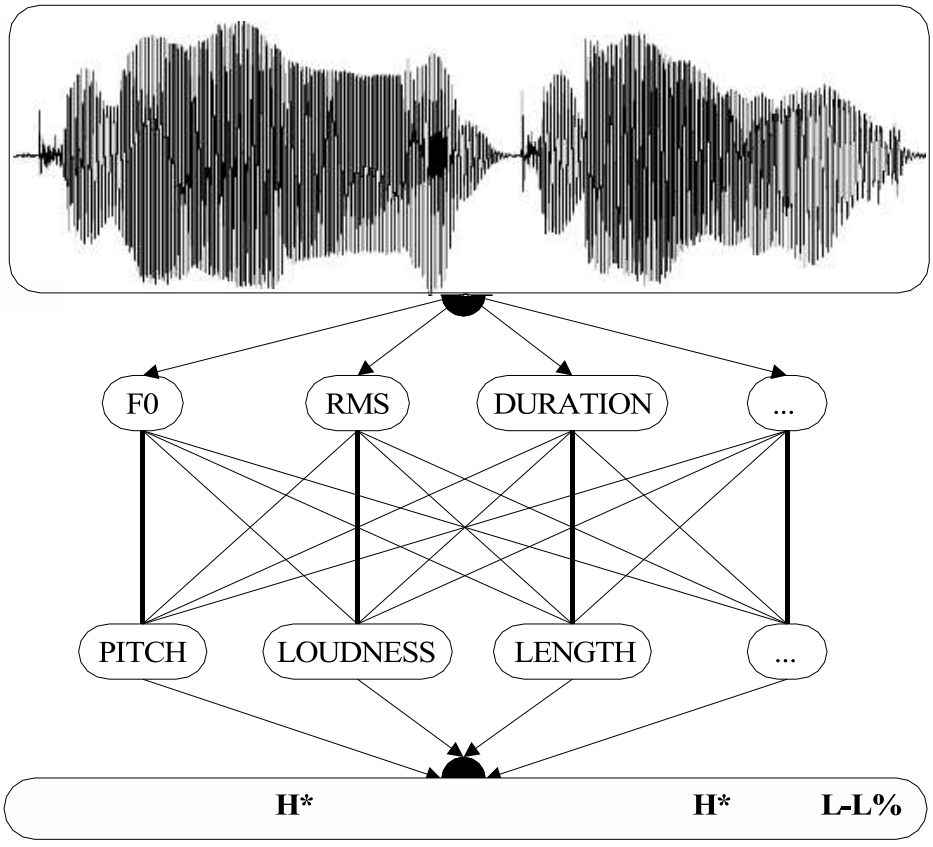
Norbert Braunschweiler

Automatic Detection of Prosodic Cues



2003

Dissertation zur Erlangung des akademischen Grades
des Doktors der Philosophie
an der Universität Konstanz
vorgelegt von

Norbert Braunschweiler

Konstanz, im Mai 2003
Überarbeitete Version, Oktober 2003

Gutachter:
PD Henning Reetz
Prof. Aditi Lahiri
Prof. Carlos Gussenhoven

Für Elena und Maxim

Contents

1 Introduction **13**

1.1 Motivation 13

1.2 Structure of the Thesis 17

2 Examples of Intonational Phenomena **19**

2.1 Offering Contour 20

2.2 Calling Contour 20

2.3 Surprise Contour 22

2.4 Focussing 22

2.5 Phrasing 26

2.6 Typological Aspects 26

3 Literature Review **31**

3.1 Discussion of Intonation Models 31

3.1.1 The Dutch School of Intonation 33

3.1.2 KIM - The Kiel Intonation Model 36

3.1.3 Fujisaki's Model 38

3.1.4 Taylor's RFC-Model 40

3.1.5 Pierrehumbert's Model 43

3.1.6 Autosegmental-Metrical Theory 48

3.1.7 Comparison of Models 50

3.2 Labeling Methods 52

3.2.1 ToBI 52

3.2.2 INTSINT 54

3.2.3 GToBI 56
3.3 Existing approaches about automatic recognition of prosodic events 60
3.3.1 Pierrehumbert 60
3.3.2 Wightman and Ostendorf 61
3.3.3 Taylor (RFC-Model) 63
3.3.4 Rapp 65
3.3.5 Ostendorf & Ross 66
3.3.6 MOMEL 68
3.3.7 Verbmobil 69
3.3.8 ToBI Lite 72
3.3.9 Other Approaches 73
3.3.10 Summary 74

4 ProsAlign - the Automatic Prosodic Aligner 77
4.1 What are the Relevant Acoustic Features? 77
4.2 Method of Parameter Assessment 80
4.3 Results 85
4.3.1 Results for Pitch Accents 87
4.3.2 Results for Boundary Tones 92
4.3.3 Conclusion 94
4.4 Phonological Mapping 94

5 Implementation of the Model 99
5.1 Faulty or Microprosodically Affected F0 Values 102
5.2 Detection of Acoustic Features 105
5.3 Acquisition of Quantitative Criteria 106
5.3.1 Parameters in the Voicing Domain 106
5.3.2 Parameters in the F0 Domain 109
5.3.3 Parameters in the RMS Domain 117
5.3.4 Summary 121
5.3.5 Results 121
5.4 Phonological Mapping 132
5.5 Rule-Based vs. HMM 133

6 Evaluation of the Program **135**

6.1 Introduction 135

6.2 First Evaluation 138

6.2.1 Method 138

6.2.2 Results 140

6.3 Second Evaluation 151

6.3.1 Method 151

6.3.2 Results 152

6.4 Discussion 157

7 Conclusions **161**

7.1 Summary of Main Findings 164

7.2 Applications for the Program 165

7.3 Future 166

Appendix **169**

A Examples of Labeled Speech Files 169

B Notes on the Computer Implementation 187

Preface

The process of information extraction from acoustic speech signals involves not only the recognition of segmental features, phonemes, syllables or words and subsequent linguistic processing, but also the recognition of prosodic events including the position of accented words, the type of pitch movement associated with them, the general trendline of pitch and also the grouping of information units, phrases or words.

The prosodic events are important conveyors of the information structure in utterances, which this work aims at unfolding for improved speech analysis and recognition. To fulfill these aims, the following tasks are done: (i) review and discussion of intonation models, (ii) development of a new approach for the automatic detection of prosodic cues, (iii) acoustic analysis of cues of prosodic events, (iv) implementation of algorithms for detecting these prosodic cues, and (v) evaluation of the new approach. Important aspects of the thesis include integration and evaluation of linguistic theory and quantitative acoustic modeling.

Acknowledgements

I would like to thank Aditi Lahiri who gave me the impetus to go into the area of experimental phonology and phonetics, and for her enduring support. She and Henning Reetz offered me the possibility to connect the worlds of physics and linguistics. Henning Reetz supported me not only during my stay at the University of Konstanz and deserves my special gratitude for his technical and intellectual support during the development of my thesis. I would also like to express my thankfulness to Carlos Gussenhoven who (probably unknowingly) helped me significantly during the design of the programs concept. Achim Kleinmann helped me always with all kinds of things regarding computers and whenever the topics to discuss seemed to be finished we could go on discussing about Linux. I am also obliged to Jennifer Fitzpatrick-Cole who supported me during our co-operation in the ILEX project. I would also like to thank my colleagues at the department of linguistics at the University of Konstanz where I started to develop the program and where the focus of interest was more related to linguistic aspects. Whereas afterwards, when I started to work at the Institute of Natural Language Processing (IMS) in Stuttgart the focus was more on the processing of speech on computers. Grzegorz Dogil offered me the possibility to work at the IMS and supported my PhD project. Bernd Möbius helped me with fruitful discussions. My colleagues at the Institute of Natural Language Processing in Stuttgart deserve thanks for their support. Martin Barbisch supported me with great enthusiasm during the transformation of my programs source code into a proper C/C++ program. Martine Grice brought me in contact with Ulrike Gut, who sent me the label data of four human labelers, which helped me a lot during the evaluation of ProsAlign.

I am indebted to Lena and Maxim for their love and the support during the development of the thesis and I hope to be more present after this thesis has been finished now. I would also like to express my gratitude to my family who always encouraged me during this PhD project and remained an oasis of relaxation and diversion. In the rare case that there are mistakes in this thesis I am the person to demand an explanation for.

Norbert Braunschweiler Konstanz, October 2003

Chapter 1

Introduction

1.1 Motivation

This study is about an approach that formulates an explicit way from continuous acoustic parameters to discrete and abstract phonological entities. The method is implemented in a computer program and uses a linguistic theory about the underlying structure of prosody in speech. The program is designed to automatically detect the position of prosodic events from acoustic speech signals. Such a program can be of great benefit for the linguist working with large acoustic databases. It enables the researcher to process unlabeled speech material automatically and systematically. The program can search for specific intonational patterns in a given language, or can test a theory about the underlying structure of prosody against the acoustic reality or the language learner can use it by seeing some visual feedback to his or her freshly acquired foreign language abilities. Furthermore the program can be used for labeling prosodic events in a spoken speech synthesis corpus and consequently improve the synthesis quality. Last but not least there are possible applications in the field of automatic speech recognition.

Prosody is used in speech communication as a supplementary knowledge source, providing information not available from the lexical meaning of the words alone. Prosodic features are variations in pitch, length, loudness and rhythm during a stretch of speech. Traditionally the term 'prosody' was used to refer to the characteristics and analyzes of verse structure. In the present study the analysis of prosody encompasses two 'worlds': on the one side is the physical world including the acoustic speech signal and its measureable entities fundamental frequency, duration, and intensity.[1] On the other side is the abstract world including per-

[1] These three entities are all physically measurable each having a unit and a fixed definition of how to extract it from a waveform (acoustic speech signal). The units are: fundamental frequency or F0 measured in Hertz [Hz], duration measured in milliseconds [ms], and intensity measured in RMS (**R**oot **M**ean **S**quare)-amplitude [Pa=Pascal] or decibel RMS-amplitude [dB_{RMS}]. See e.g. Reetz (1999, p. 19 ff), for more detailed information about these parameters.

ceived entities of pitch, length, and loudness as well as linguistic representations that are assumed to play a crucial role in the process of speech understanding. Both 'worlds' are connected in speech recognition and understanding. Utterances are expressed with variations in frequency, duration and loudness and these units are the conveyors of informations, ideas, instructions, etc. However, to become information the physical parameters have to be interpreted by a listener and it is a common observation that obviously different acoustic signals can be interpreted by listeners as conveying the same information. For instance the word "information" uttered by a male and a female speaker in the same context may show clearly different individual acoustic properties like segment durations, energy contours, F0 movements, etc. but both realizations are usually easily interpretable by human listeners as conveying the same "information". Abstraction from measureable acoustic parameters towards meaningful units is a process that is not easily manageable by machines. This everyday experience is still a controversial subject in the field of linguistics and automatic speech recognition. The present study focuses on a part of these processes, namely the extraction of prosodic information from the acoustic signal (cf. figure 1.1). The mentioned parameters are most important for the perception of prosodic events, but additional parameters may contribute as well as is symbolically expressed by the unfilled boxes in figure 1.1. One of the additional parameters could be for instance the formant values which are the most dominant acoustic correlates of perceived phoneme quality.

This study uses explicitly the term 'prosodic cues' in its title to state that not only variations in F0 are taken into consideration, but also variations in duration and intensity. Although the term 'intonation' is often used interchangeably with 'prosody',[2] it is usually used to refer solely to variations in pitch and subsequently only to variations in F0. Here both terms will be used interchangeably but when terminological differences appear they will be mentioned.

Intonation is used in communication to express differences of expressive meaning (e.g. happiness, surprise, anger). It is also very important for the naturalness of language, which is of course most obvious in speech synthesis systems.[3] Beside the latter aspects, intonation serves a grammatical function distinguishing one type of sentence from another. Thus, a phrase like *Hundred Euro* said from a cashier behind the counter when one has to pay for something that is worth the price like a DVD-player or the newest book about the latest linguistic model usually begins with a high or medium pitch and ends with a lower one (i.e. falling melody) is a simple request, whereas *Hundred Euro?* said as response to the same request but for paying something whose value is far away from the price demanded, like a bag of popcorn or two lollypops, will be usually expressed with a rising melody (ending in a high pitch) or even a rise-fall-rise melody and increased emphasis,

[2] See e.g. Hirst & Di Christo (1998, p. 3 ff) for a more detailed discussion of this terminological problem.

[3] The present study focuses on the automatic analysis of prosody and therefore does not explicitly deal with aspects of prosody for speech synthesis purposes.

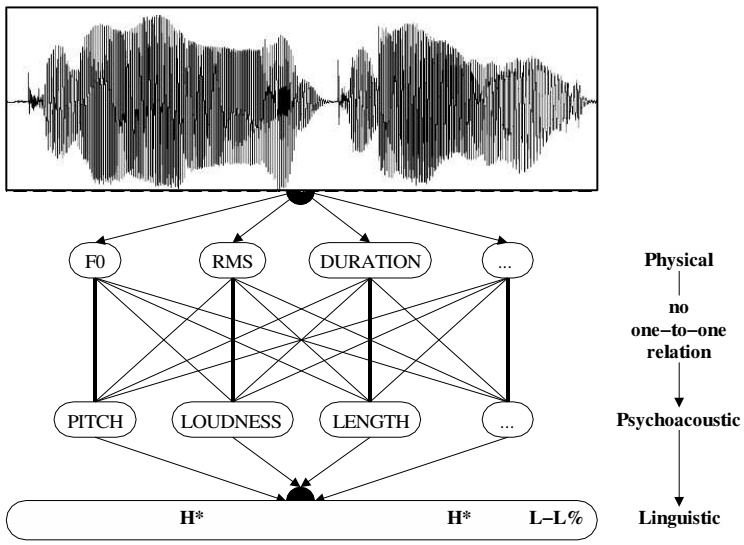


Figure 1.1: Depiction of the physical and perceptual levels in the process of intonation perception. From the acoustic speech signal acoustic features are extracted and related to perceptual dimensions. There is no one-to-one relation between the physical level and the perceived entities. Most dominant relations are marked with thicker lines. Unfilled boxes indicate additional parameters not already depicted.

and indicates a surprise question (see also 2.3 and Ladd 1996, p. 43 ff for the discussion of a rising-falling-rising tune). Additionally these melodies may be used for different purposes in different languages, that is, they are language dependent.

This example shows that the same sentence can be expressed with different *intonational tunes*.[4] In phonology a tune is usually characterized as a structured sequence of abstract intonation labels and is associated with a functional aspect.[5] Each of these tunes could have consequential influences on the interpretation of the sentence. The other way around the same tune can be overlaid on many different sentences (as will be shown in 2.1). Therefore intonation conveys additional information to the selection of words and their lexical meaning, to mark communicative purposes, like asking a question, emphasizing a specific word or a part within a sentence, structuring the speech in specific ways, or simply sounding funny, humorous, depressed, etc. One of the tasks in linguistic modeling is to set up a sat-

[4]When terms are introduced the first time they are written in italics.

[5]The labels are called 'abstract' because they are not exactly defined in terms of concrete quantitative limits but are thought of as covering a wide range of acoustic events that build a distinct perceptual class from another abstract label. A specific notational system that describes the structure of tunes is presented in chapter 3.2.1. "[...] tunes are linguistic entities, which have independent identity from the text. Tunes and texts cooccur because tunes are lined up with texts by linguistic rules." (Pierrehumbert, 1980, p. 19).

isfying description of a specific subset of intonational phenomena, namely those which do not express some sort of *paralinguistic interpretation*.[6]

A linguistic model should be able to explain explicitly the underlying processes and structures in the recognition process. Therefore a purely acoustically based analysis can only give very limited insights. This is reflected by the problems of automatic speech recognition systems to deal with acoustic variation without including a model of the underlying structure of a given language. With respect to the automatic recognition of prosodic patterns this means that a purely acoustically based analysis system could achieve only a limited recognition of principally different prosodic patterns. In this thesis the working hypothesis is, **that the acoustic analysis is the 'igniting device' for a general process 'prosody recognition'**.

The whole process involves crucially the formative influence of a predefined or acquired linguistic structure on the acoustic continuum. One of the aims in this study is to uncover the rules of this process and to formalize them. This faces us with a number of problems, because we have to deal with strong variation in the acoustic parameters, where the source of variation is often unclear or results from a complex interaction of many factors. The approach presented here tries to take the different sources of variation into account and to handle them in an integrated approach of automatic detection of prosodic events. It has to be stated, however, that this is only a part of the whole process of speech recognition and understanding. A complete system would have to identify the individual segments, syllables, and words as well. Often this segment detection was the only analysis strategy in former (and still in most of the current) automatic speech recognition systems and larger units ('supra-segmentals') had not been taken into account. However, prosody is incorporated into automatic speech recognition systems (e.g., Hess et al. 1997; Batliner et al. 2001b).

What is meant by the title of the thesis: "Automatic detection of prosodic cues"? First of all what is presented is an "automatic" procedure, that means there is no hand labeling involved. During the development of the algorithm manually labeled data was used only for the acquisition of selection criteria. All steps in the process are executed in a computer program. The input to the program is a speech signal and the output is a set of labels with information about the type of prosodic event and where it appears in time in the speech signal (see figure 1.2). The procedure involves no segmentation of the speech signal into words, syllables or phonemes before the abstract prosodic entities are determined. It is solely the (sometimes complex) amalgam of the above mentioned acoustic parameters that is taken into consideration as an initiation of the search for adequate prosodic entities. Both, *bottom-up* (from acoustic-to-phonological entities) and *top-down* (from phonological-to-acoustic entities) processes are involved to determine the abstract

[6] Paralinguistic intonational phenomena are differences of sex, age, social status, sadness, etc. These distinction is drawn to focus on the underlying linguistic structure and not on speaker individual or task specific specialties. However, a distinction is not always clear cut. See also the discussion of this subject in Ladd (1996, p. 33 ff).

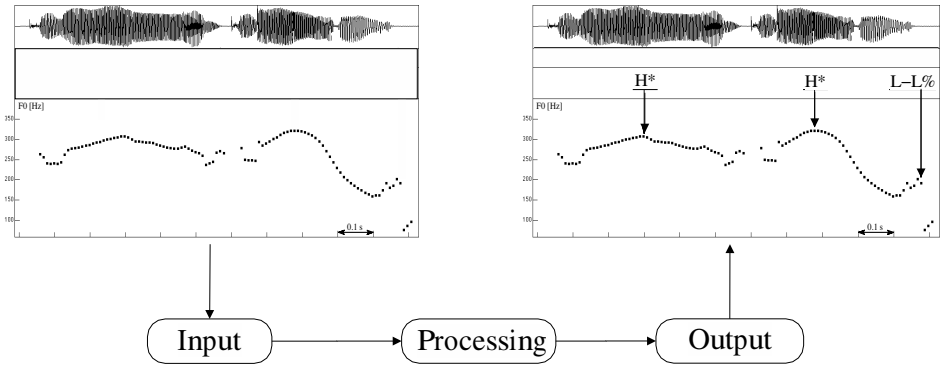


Figure 1.2: Illustration of input and output of the computer program presented in this thesis. The input is a speech signal and the output is a label file with information about the type of prosodic event (pitch accents and boundary tones according to the ToBI model, see section 3.2.1) and where it appears in time.

entities. The bottom-up procedure is the search for acoustic cues given from the course of F0 and RMS amplitude;[7] the top-down procedure is represented by the mapping logic from abstract labels to acoustic cues.

1.2 Structure of the Thesis

The basic structure of this thesis consists of a presentation of some intonational phenomena, a literature review, a description of the method chosen for automatic detection of prosodic events, an evaluation of the method and finally a discussion. In more detail: In chapter 2 some intonational phenomena like typical contours or text-to-tune alignment are laid out. Chapter 3 presents a literature review of the most influential theories about intonational structure or F0 modeling: the **IPO** theory of intonational structure ('t Hart et al. 1990), **Fujisaki's** F0 model (Fujisaki & Hirose 1982; Fujisaki 1983), the **K**ieler **I**ntonation **M**odel (**KIM**; Kohler 1991), Taylor's **RFC**-model (**R**ise/**F**all/**C**onnection; Taylor 1994), and **Pierrehumbert's** theory (Pierrehumbert 1980; Beckman & Pierrehumbert 1986). A sketch of the basic principles of the "autosegmental-metrical theory" (see Ladd 1996 and section 3.1.6) is given afterwards. The chapter also compares the presented models and presents the phonological modeling of intonation in German in more detail. Further two labeling instructions are presented in this chapter and finally some approaches about the automatic detection of prosodic events are introduced and discussed.

[7] Duration, as mentioned on page 14 is not measured directly but influences the steepness of F0 and RMS curves, that make up the 'course of F0 and RMS amplitude'

In chapter 4 the outline of the automatic **pro**sodic **align**er (ProsAlign) is introduced. Chapter 5 describes the implementation of the model in a computer program and chapter 6 its evaluation. Finally chapter 7 summarizes the findings in this work and discusses future directions. Terminological questions will be dealt with when the term in question first appears and will be explained in footnotes.

Chapter 2

Examples of Intonational Phenomena

It is common knowledge that the way something is said can be just as important in conveying a message as the words used to say it. In order to present some examples of the latter and to give an insight in the field of work the following chapter will present some of the intonational phenomena in German and also a sketch of typological aspects of intonation, since other languages might use different contours for the same type of sentences. According to Helfrich (1985) there are three functions of intonation that modify an utterances meaning: (i) marking of sentence type, (ii) focussing, and (iii) disambiguation.[1] For the first case examples of *offering contours*, *calling contours*, *surprise contours* as well as typical contours from *declarative* and *interrogative* sentences are presented. Some of these examples show the effect of the overlay of one and the same intonation contour on different text material. In turn, other examples show how one and the same text material is aligned with different intonation contours. *Focussing* is illustrated by a question-answer example. *Phrasing* is outlined with an example of the same text that results in two totally different meanings according to the different subdivision into prosodically coherent units. Finally some language universal aspects of intonation are addressed. This chapter should lay the ground for what sort of abstract information should be extracted by the proposed algorithm in chapter 4.

The following illustrations show waveforms from speech files and time aligned F0 contours as extracted by the ESPS/waves pitch tracker *get_f0* (version 1.14).[2] The procedure to extract the F0 contour from a given waveform can be roughly characterized as the detection of the more or less periodically repeating glottal pulses in its voiced segments. However, this is only one method of extracting F0, there are

[1] Crystal (1995, p. 249) describes six functions of intonation: emotional, grammatical, informational, textual, psychological and indexical, see also the definitions of intonation in (Rossi, 2000, Section 2.4.3).

[2] See more information regarding the *get_f0* program on page 99.

also articulatory based procedures that measure vocal fold vibration with a *laryngograph* by attaching electrodes to the neck of a speaker and also auditory perception as it was more often applied in former years. F0 is defined as the number of glottal pulses per second (= Hz). The periodicity of human sound signals is known to be not perfectly periodic therefore it is also called 'quasi-periodic' (see e.g. Hess 1983). Problematic aspects of the automatic F0 extraction will be addressed in chapter 5.1.[3]

2.1 Offering Contour

The so called *offering contour* is usually used when somebody wants to offer something to somebody else as in the questions given below.

(a) Willst du Kaffee? *Do you want coffee?*

(b) Willst du noch mehr Vanilleeis mit Schlagsahne? *Do you want some more vanilla icecream with whipped cream?*

The offering contour is intonationally realized as a fairly constant beginning up to a rise at the end. The F0 contours from the offering contour examples (see figure 2.1) show that there is different sized text material aligned with what counts phonologically as one and the same contour.

2.2 Calling Contour

The so called *calling contour* or *vocative chant*[4] is usually used for calling somebody who is not in the immediate vicinity of the caller. Therefore the caller normally raises his/her voice to get the attention of the addressee. The contour is characterized by a fall from a high level in the *nuclear accent*[5] to a mid level at the end. Figure 2.2 shows the waveforms and F0 contours of two names that are pronounced with a calling contour:

(a) Mar'ia *Mar'ia*

(b) Heide'linde *Heide'linde*

[3] For an overview of approaches of F0 extraction see Hess (1983) and Reetz (1996).

[4] See for instance Gibbon (1976, p. 274-287); Ladd (1978); Féry (1993, p. 96 ff); and Ladd (1996, p. 88, 136 ff) for more detailed information about this type of contour.

[5] The nuclear accent is usually the last accent in the intonation phrase. Originally the term nuclear tone was introduced by the British School of intonation (see Palmer 1922; O'Connor & Arnold 1961; Halliday 1967; Crystal 1969) description and referred to a typical pitch movement like for instance a 'rise-fall-rise'. This way of intonation description helps to see "how the tune in question is applied to texts with varying numbers of syllables and different stress patterns" (Ladd, 1996, p. 44).

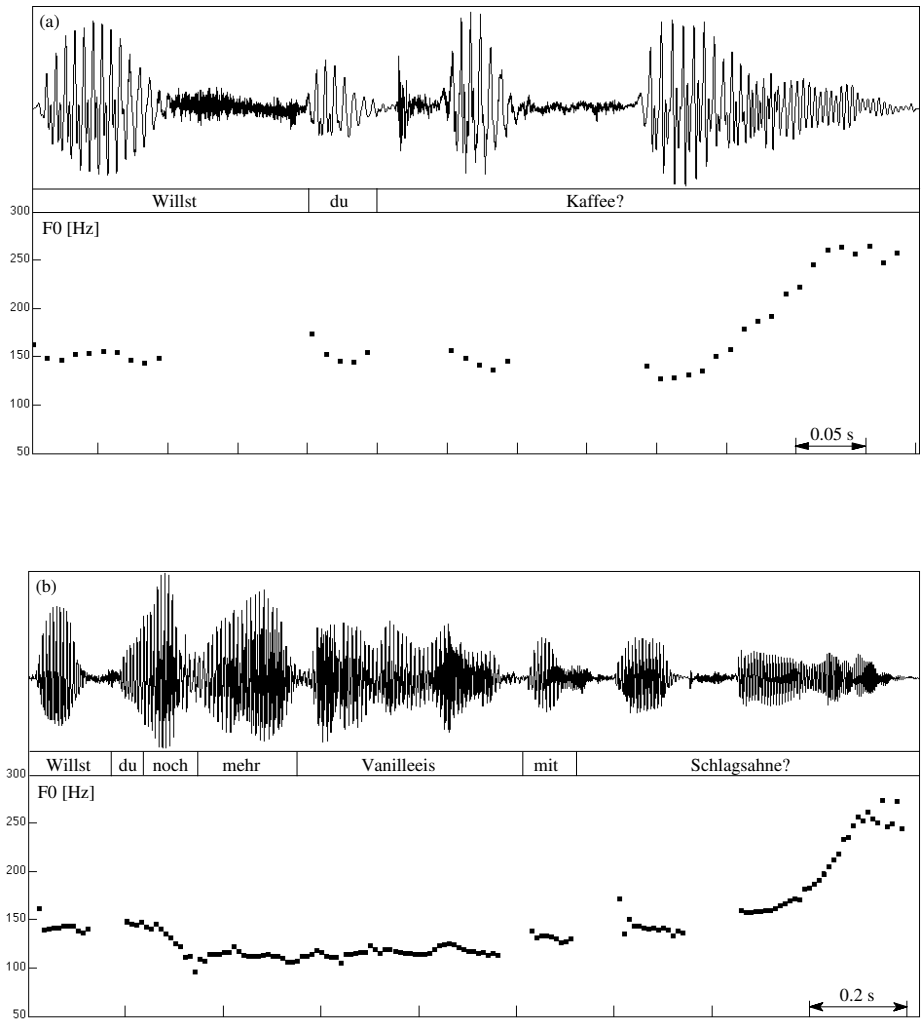


Figure 2.1: Two different sentences having the same underlying offering contour. Both sentences were uttered by the same male speaker. Note the different time scales in the two views.

Féry mentions that a "crucial property of a stylized contour is that the syllables involved are lengthened, the last (full) one being the longest." (Féry, 1993, p. 101). Among the stylized contours mentioned by Féry the calling contour is the most typical one.

2.3 Surprise Contour

The surprise contour is usually used when somebody is very astonished about something. This type of contour is typically realized with a sharp rise at the end of the utterance preceded by a fairly low part and optional accents before (see figure 2.3 a).

(a) Das ist Maria? *That is Maria?*

The same text of the surprise question spoken as a declarative is also depicted for comparison (see figure 2.3 b).

(b) Das ist Maria. *That is Maria.*

The declarative is realized with a falling contour basically. These two examples show how the same text may be altered in its information content by the intonation contour.

2.4 Focussing[6]

When sentences are analysed by linguists they may be separated with regard to the information that is known by the speakers, and that which is at the midpoint (or 'focus') of their conversation. Therefore focus is opposed to presupposed subjects. For example, when one wants to emphasize a certain contrast it is possible to use prosodic means to put the object of emphasis in focus.[7] Focus is often not predictable from the syntactic structure and strongly influences the meaning of a sentence. Below is an example of a short conversation:

(a) Willst du zwei Eier haben? *Would you like to have two eggs?*
Nein, ich möchte EIN Ei haben! *No, I would like to have ONE egg!*

[6] There is a considerable amount of literature dealing with aspects of focussing and intonation, for instance especially concerning German Féry (1993) and Uhmann (1987), and more general in Ladd (1996, chapter 5).

[7] Of course there are other means to signal focalization. Gibbon (1998, p. 89) mentions the three possible means for German: focus particles, word order, and accentuation.

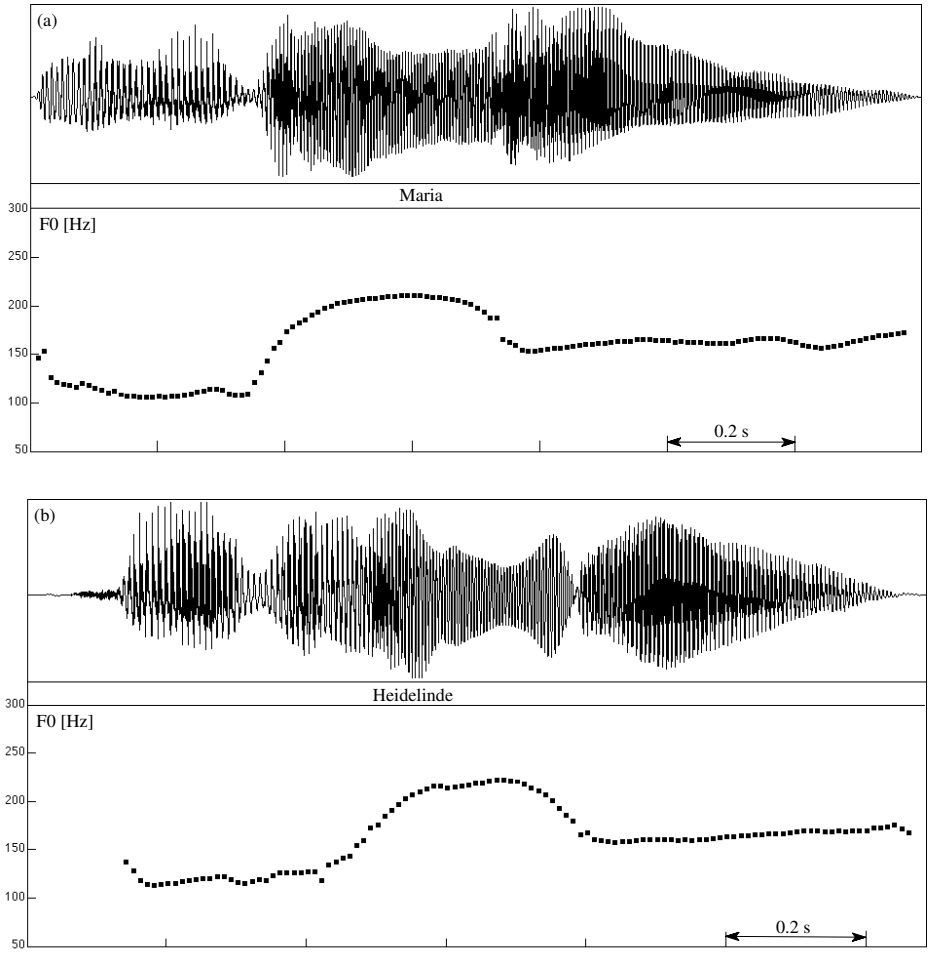


Figure 2.2: Two examples of calling contours. In (a) the calling contour is applied to the name "Mar'ia", in (b) to the name "Heide'linde". Both cases can be thought of somebody calling the named person who is not in their immediate vicinity. The sentences were uttered by the same male speaker. Note the different amount of segmental material for the same contour.

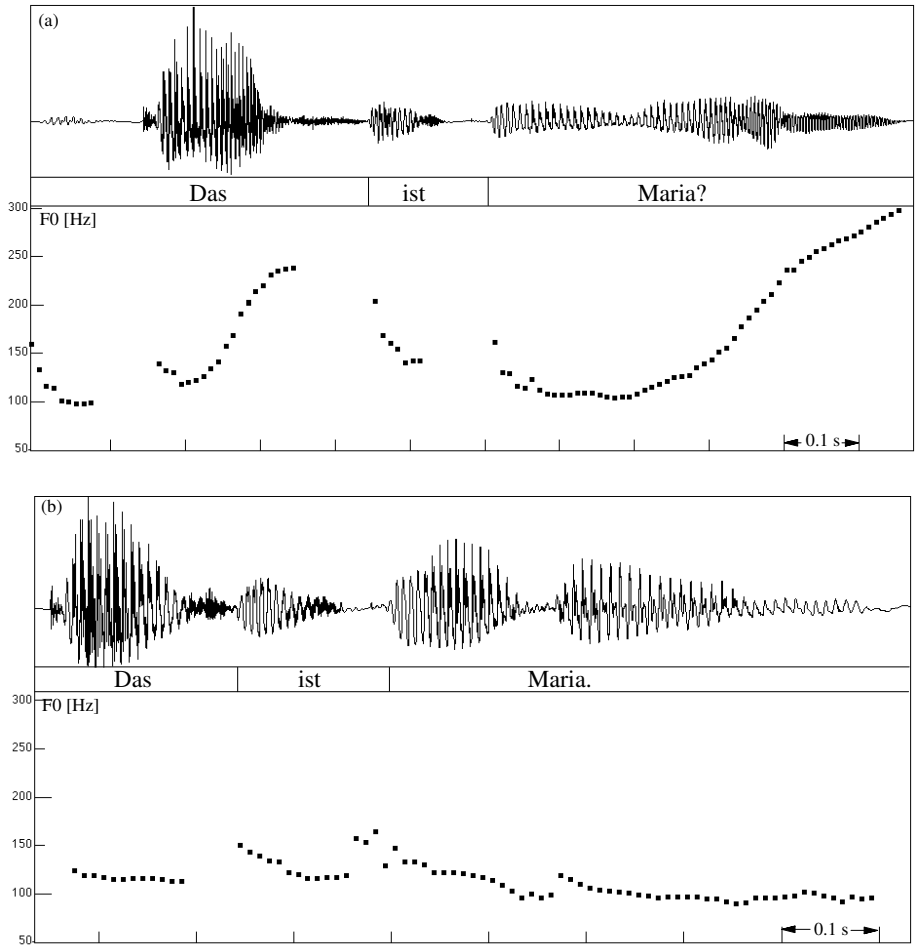


Figure 2.3: Example for a surprise contour (a) and the same sentence said as declarative (b). Both sentences were uttered by the same male speaker.

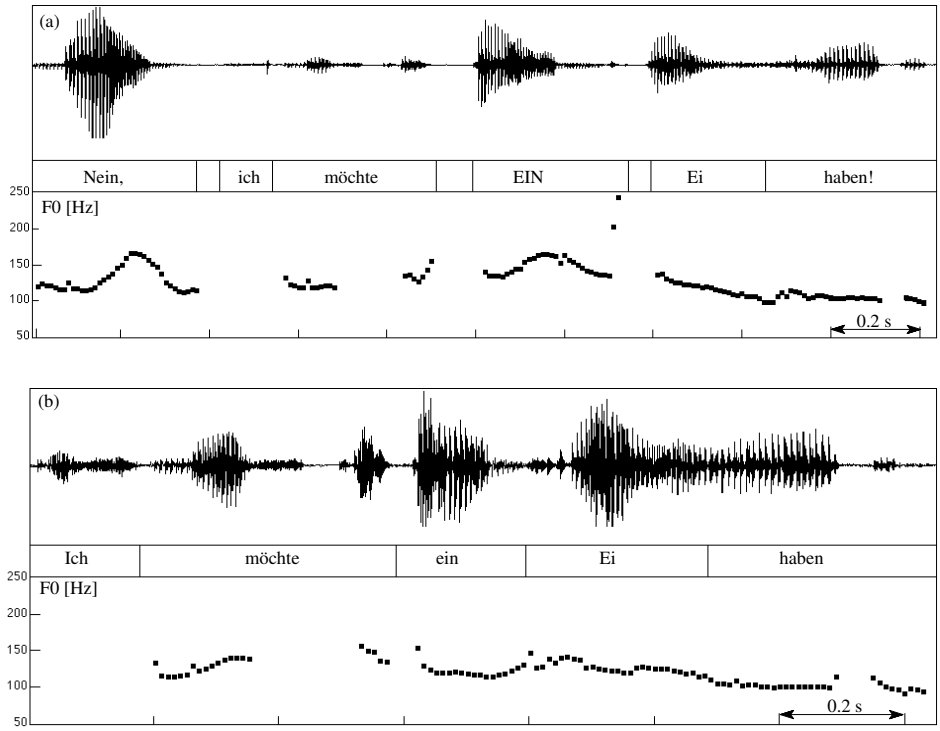


Figure 2.4: Illustration of a focus on "EIN" *ONE* in the sentence: "Nein, ich möchte EIN Ei haben!" *No, I would like to have ONE egg!* (a) and the same sentence without focus on "ein" *one* (b). Both sentences were uttered by the same male speaker.

In the answer phrase the addressee expresses his wish to get only one egg by placing a focus on "ein" *one* (see figure 2.4 a). When the contrast is pronounced really strong there are short pauses before and after the focused word and a pitch accent on the word itself. However, in other (less strongly pronounced) cases the pauses might not be present. A focus may be either broad or narrow (cf. Ladd 1980) in the first case being not limited to a specific domain whereas in the second case the focus domain is limited to a smaller constituent usually a word bearing the accent. Féry (1993, p. 71) states that native speakers of German were not able to differentiate between a broad and a narrow focus reading of the same sentence. As comparison the same sentence said without focus on "ein" *one* is illustrated in figure 2.4 b.

(b) Ich möchte ein Ei haben. *I would like to have an egg.*

The unfocused "ein" *one* is realized clearly differently as the focused one. There are no pauses before and after the unfocused word, and no pitch accent on the word itself.

2.5 Phrasing

Sometimes prosodic means are used to disambiguate a sentence that can be parsed in two different syntactic structures by separate phrasings as in the example given below (following a similar example from Helfrich 1985, p. 17):

(a) (“Hoeness”,) (sagte Daum,) (“wird nie gewinnen”). *“Hoeness”, said Daum, “will never win”.*

(b) (Hoeness sagte:) (“Daum wird nie gewinnen”). *Hoeness said: “Daum will never win”.*

The illustration of the F0 contours in figure 2.5 show the different phrasings. The inspection of the waveforms and F0 contours shows that the distinctive phrasings are realized on the one side by different placements of pauses: in (a) a pause follows the utterance parts “Hoeness” *Hoeness* and “sagte Daum” *said Daum*, in (b) there is only one pause namely after “Hoeness sagte” *Hoeness said*. On the other hand the two sentences are separated by varied accentuations (for instance, the different F0 movements and amplitude modulations on “Hoeness” *Hoeness* in (a) and (b)), continuation rises[8] at the end of intonation phrases (for instance at the end of “sagte Daum” *said Daum* in (a)), and different word durations (for instance “sagte” *said* in (a) is shorter than in (b)).

In his extensive work about rules for German sentence intonation Bierwisch formulates explicit rules of how to derive phrasing units from the surface syntactic structure. Phrasing units which are relevant for the intonation do not coincide with the syntactic constituents therefore he proposes special boundary symbols. The placement of these boundary symbols is partly determined by syntactic structure but not identical to it (cf. Bierwisch 1966, p. 106 ff).[9]

2.6 Typological Aspects

The last sections presented examples of intonation contours regarding offering, calling, surprise, focussing, and phrasing that showed some of the functions intonation is used for in German. However, intonation patterns may also be different across languages (see e.g. Ladd 1996, Ch. 4). Comrie (1984, p. 17) mentions that “English with Russian intonation sounds unfriendly, even rude or threatening, to the native speaker of English; Russian with an English intonation sounds

[8] Continuation rises (also called “progredient” intonation, cf. Féry 1993 and Gibbon 1998, p. 88) are F0 rises or perpetuations of F0 at a mid level that are used at the end of phrases to signal that the speaker has not yet finished his speech but wants to continue with something related.

[9] Bierwisch also deals with the question how sentence accent can be determined on basis of the syntactic structure in German (cf. also Kiparsky 1966).

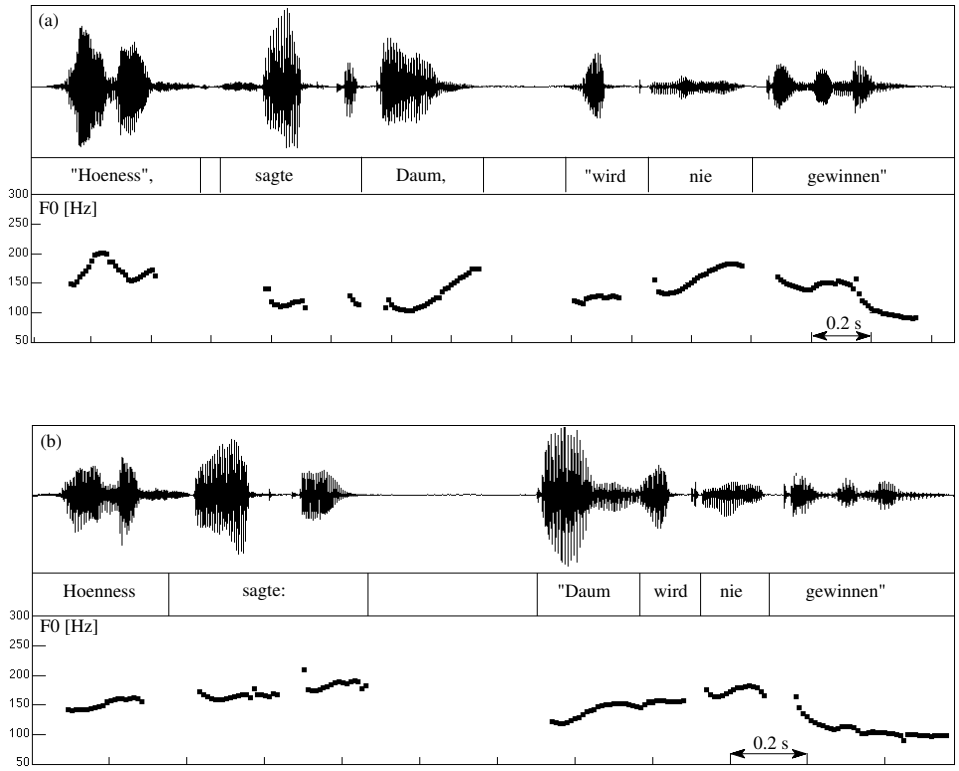


Figure 2.5: Two different phrasings of the word sequence ≪Hoeness sagte Daum wird nie gewinnen≫ Hoeness said Daum will never win.
(a) ("Hoeness",) (sagte Daum,) ("wird nie gewinnen".) *"Hoeness," said Daum, "will never win."*
(b) (Hoeness sagte:) ("Daum wird nie gewinnen") *Hoeness said: "Daum will never win."*

affected or hypocritical to the native speaker of Russian." Ladd (1996, p. 115) presents another example, namely "[...] that many Hungarian questions sound like emphatic statements to native speakers of English [...]". Regarding typological aspects of intonation Ladd (1996) refers to Bolinger's (Bolinger 1978, 1986, 1989) work: "Intonation, according to Bolinger, has direct links to the prelinguistic use of pitch to signal emotion. Broadly speaking, high or rising pitch signals interest, arousal, and incompleteness, while low or falling pitch signals absence of interest and hence finality and rest. This fundamental opposition between high and low (or up and down) is clearly seen in the use of pitch range for obviously emotional expression - raised voice for active emotions such as anger or surprise and lowered voice for boredom, sadness, and the like." (Ladd, 1996, p. 113-114). Ladd continues to summarize 3 grammaticized main usages of intonation across the languages of the world:

"1 the tendency of pitch to drop at the end of an utterance, and to rise (or at least not to drop) at major breaks where the utterance remains incomplete;

2 the use of higher pitch in questions, since in questions the speaker expresses interest, and since the exchange is incomplete until the addressee answers;

3 the use of local pitch peaks (e.g. pitch accents) on words of special importance or newsworthiness in an utterance." Ladd (1996, p. 114)

The "Universals Archive"[10], a collection of language universals which is searchable online lists 10 entries out of a total of about 2000 entries when searching for the keyword "intonation". Among them is a statement made by Bolinger (1978, p. 472): "Terminal intonations are almost universally low or falling for finality and assertion, and high or rising for the opposite, including yes-no questions."[11] (cf. entry number 1003 in the Universals Archive). In Fitzpatrick-Cole (1999) a sketch for a typology of intonation is presented. Here it is also mentioned "that Bolinger's [...] "Universalist" theory of intonation is steadily losing ground to phonological theories of intonation [...]" Fitzpatrick-Cole (1999, p. 941). In a paper about "Bengali Intonational Phonology" Hayes & Lahiri argue against a view held by Bolinger (1972) "that all phrasal stress is non-phonological in nature, reflecting only semantic factors." (Hayes & Lahiri, 1991, p. 48) and present a phonological rule of phrasal stress assignment which applies in neutral focus contexts.

[10] Cf. http://ling.uni-konstanz.de/pages/proj/sprachbau.htm and Plank & Filimonova (2000).

[11] One of the rare counterexamples for the terminal rising intonation in questions across languages is mentioned by Li & Thompson: "The opposite of marking questions with a rising intonation is found in Chitimacha, an American Indian language of Louisiana, where declarative sentences have a rising intonation and questions have a falling intonation." (Li & Thompson, 1984, p. 60)

Gussenhoven (2002) expresses the question whether intonational meaning is universal or language specific. He argues for the position "that both the universal and the language-specific perspectives are true, simultaneously, for any language, but that the universal part is exercised in the phonetic implementation, while the language-specific meaning is located in the intonational morphology and phonology." (Gussenhoven, 2002, p. 47).

For the automatic extraction of prosodic cues the actual use of a specific contour for a type of sentence is irrelevant, as long as the phonological categories used by the algorithm (c.f. chapter 4.4 and 6.2.2.2), are appropriate for the language. Though, it has to be mentioned that the same sentence type can be realized differently in different languages.

After this short overview of some intonational phenomena in German and the reflection of typological aspects of intonation the next chapter will review the ways of prosodic modelling in the literature.

Chapter 3

Literature Review

One of the basic questions in linguistic research was and still is how to set up an adequate model of prosodic structure. What is an adequate model that is applicable to the full range of languages and provides interfaces to other linguistic modules like syntax, semantics and phonology? How to abstract away from the acoustic detail to cover the basic intonation contours occurring in a given language? How to describe intonational phenomena with a structured set of linguistically meaningful units, that is how to label acoustic speech data with linguistically meaningful units? Since most of these questions are central for the approach here the following chapter reviews existing intonation models in the first part and describes explicitly three methods (ToBI, INTSINT, and GToBI) of prosodic transcription (also called labeling instructions) in the second part.

Since the automatic detection procedure developed in this thesis was mainly applied to German, data its intonational description and modelling will be introduced in more detail in the third part of this chapter. The last part of this chapter will present existing approaches about automatic detection of prosodic events and will end with a comparison of them.

3.1 Discussion of Intonation Models

This chapter presents some of the existing models that describe intonational phenomena. The main interest here to look at these models is focused around their usefulness for automatic detection of intonational cues. There are, however, a number of theoretical implications that are worth discussing. When looking at the relevant literature in this field one is overwhelmed by the number of different terminologies as well as the number of problematic issues within and between the models. Even in more recent textbooks (e.g., Ladd 1996; Hirst & Di Christo 1998) there is neither a commonly accepted model of intonation nor a generally accepted standard

for describing intonation. However, efforts are made to set up a standard for transcription (e.g. INTSINT = **IN**ternational **T**ranscription **S**ystem for **INT**onation in Hirst & Di Christo 1998, see also the discussion later in this chapter) that are trying to fill this gap, but see the critique of this approach in chapter 3.2.2.

Of course, each individual transcription system has its own value, and it is, although with some limitations, possible to transform one transcription into another, but (not only) for reasons of theoretical wellformedness it would be better to have such a more generally accepted transcription system. Such a system could be in the sense what Hirst & Di Christo (1998) called a "Third Generation" model of intonation "which would go beyond single language descriptions (first generation) and multi-language descriptions (second generation) by defining a number of independent levels of representation determined by more general linguistic principles" (Hirst & Di Christo, 1998, p. 43).

Additionally, a number of researchers have adopted more or less the basic principles of the ToBI transcription system (for **To**ne and **B**reak **I**ndices, cf. Beckman & Ayers 1997) that was originally developed for describing the intonation patterns and other aspects of the prosody of English to other languages (e.g., Bengali: Hayes & Lahiri 1991; Dutch: Gussenhoven et al. 1999; German: Grice & Benzmüller 1995; Greek: Arvaniti & Baltazani, to appear; Japanese: Venditti 1995; Korean: Beckman & Jun 1996).

A model of the intonational structure of a given language has to show its descriptive force by its ability

- to describe intonation contours showing clearly different communicative force with distinct phonological descriptions,
- to describe acoustically very different intonation contours that are only variants of one and the same underlying contour with one and the same underlying phonological description, and
- to cover the whole range of intonational phenomena in a given language with the proposed inventory of categories.

Existing intonation models can be roughly divided into two basic classes, viz. the ones that are looking at intonation first from the acoustic-phonetic side ("bottom-up"; e.g. the Dutch school, summarized in 't Hart et al. 1990; KIM, (Kohler, 1991); Fujisaki, (Fujisaki & Hirose, 1982); Taylor, (Taylor, 1994)) and the others that are first looking at the phonological side ("top-down"; Pierrehumbert 1980; Ladd 1996). This does not imply that the models do not take into account results from the other side, but in dependence on their starting point the models differ in their explanatory force and usefulness for the purposes of automatic detection of prosodic events. The models belonging to the "bottom-up" side were sometimes more focused in finding rules and an inventory of intonation patterns for text-to-speech

synthesis, whereas the approaches belonging to the "top-down" side grew out of theoretical problems in phonology (see Ladd 1996, p. 42). Ladd subsumes the latter group as models that belong to the "autosegmental-metrical (AM)" theory.[1] The autosegmental-metrical theory "adopts the phonological goal of being able to characterize contours adequately in terms of a string of categorically distinct elements, and the phonetic goal of providing a mapping from phonological elements to continuous acoustic parameters" (Ladd, 1996, p. 42).

The subsequent chapter reviews the following models: (1) the Dutch School, (2) the Kiel model of intonation (KIM = "**K**ieler **I**ntonations **M**odell"), (3) Fujisaki's articulatory based model, (4) Taylor's RFC-model (**R**ise/**F**all/**C**onnection), (5) Pierrehumbert's model as an exemplary case of what Ladd (1996) calls the *autosegmental-metrical* approaches. Afterwards the basic principles of the latter ones are sketched and finally the models are compared.

3.1.1 The Dutch School of Intonation

The so called "Dutch School of intonation" is an attempt started in the early sixties at the "**I**nstituut voor **P**erceptie **O**nderzoek" (IPO) and is summarized 1990 in 't Hart et al. (1990) to describe intonation from a perception point of view. The researchers started with the description of Dutch intonation and applications of their model to other languages have been also conducted (Russian: Keijsper 1983; German: Adriaens 1991). The basic assumption underlying their research is as follows:

> "[...] that only those F0 changes would be regarded as possible candidates for a descriptive model of pitch for which a link could be established with commands to the vocal-cord mechanism, which as such are under the speaker's control" ('t Hart et al., 1990, p. 186).

Although this is an articulatory based assumption and the authors are stating explicitly that physiological measurements should be made, they also state that "such a method has a number of unattractive aspects" (ibid, p. 39). The number of speakers would be restricted to those who want to volunteer for such experiments, also the authors doubt that under this experimental circumstances spontaneous speech could be recorded. However, the articulatory based assumption is said to have a consequence, namely "that the involuntary fluctuations do not make an essential contribution to the perception of the speech melody: their omission [...] should

[1] Autosegmental phonology was originally invoked by Goldsmith (1976) and contrasts with strictly segmental theories of phonology. In traditional segmental phonology a representation consists of a linear arrangement of segments. Whereas in autosegmental phonology a representation consists of several 'tiers', each tier including a linear arrangement of elements which are linked to each other by association lines. See e.g. Goldsmith (1976), Gussenhoven & Jacobs (1998).

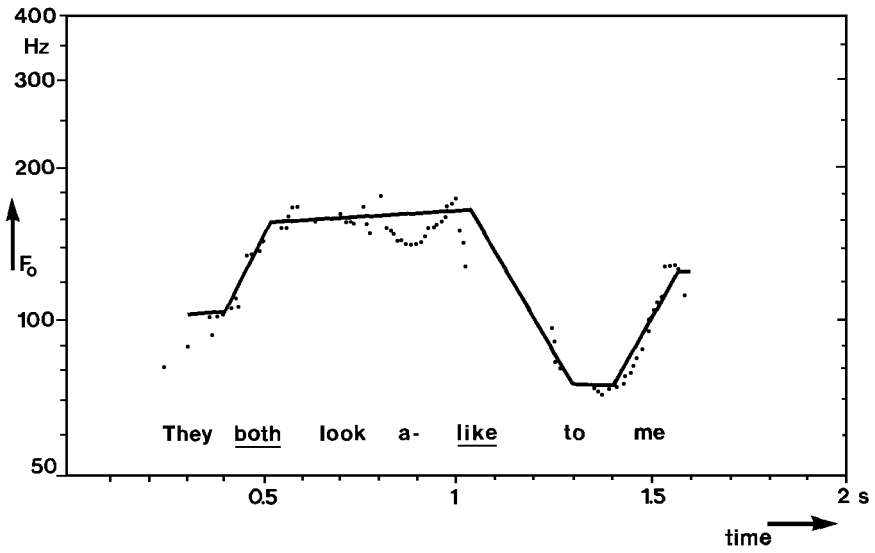


Figure 3.1: Illustration of "close-copy" stylization from 't Hart et al. 1990, p. 43). The F0 values are depicted on a logarithmic scale.

not bring about any substantial change in the perceived speech melody" ('t Hart et al. 1990, p. 40). This view expresses the observation that not all details in the F0 contour are relevant in perception. Therefore the central question in the IPO theory was: what are the perceptually relevant pitch movements? The strategy to find these relevant pitch movements is based on two steps: *stylization* and *standardization*. In the process of *stylization* a F0 contour is taken and straight lines are drawn to fit to the original contour ("close-copy"; see figure 3.1). The stylized contour has to be perceptually equal to the original one. This is tested by re-syntheses of the stylized contour and comparing it to the original contour. However, since the quality of speech synthesis was at that time not as advanced as it is nowadays it does not seem to be a convincing procedure. The process of *standardization* (see figure 3.2) involves the adaptation of the stylized contour to a grid of three continuously decreasing lines (L(ow), M(id), H(igh); representing the declination effect[2] under the criterion of perceptual equivalence to the close-copy representation.

This approach describes intonation contours with series of straight lines falling between the three declination lines. Pitch accents are represented by rises and falls between these declination lines. The procedure includes a clear reduction of information at each stage from the continuously varying F0 contour up to the standardized intonation patterns.

[2]Declination is usually understood as the slight fall of pitch during the beginning and end of an intonation phrase. It is however, far from being uncontroversial. The question under dispute here is whether it is an actively controlled process or resulting from other processes (see e.g., Hirst & Di Christo 1998, p. 21)

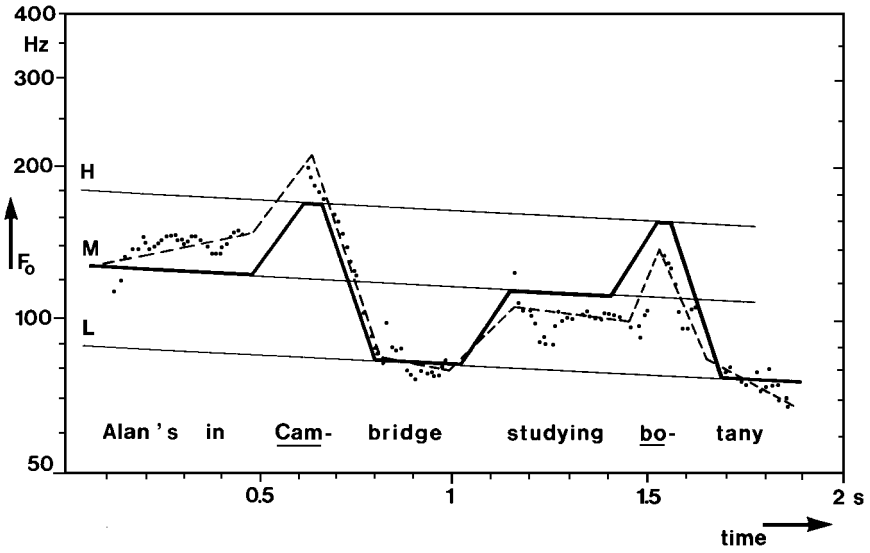


Figure 3.2: Illustration of "standardization" in the IPO approach (from 't Hart et al. 1990, p. 49). The F0 values are depicted on a logarithmic scale.

In the book published in 1990 ('t Hart, Collier and Cohen) the authors are criticizing the 'levels' approach (under which they subsume the approaches by Ladd 1983 and Pierrehumbert 1980 and others) several times. The main line of criticism is that the author reject the view that "the speaker primarily intends to hit a particular pitch level and that the resulting movements are only the physiologically unavoidable transitions between any two basic levels" ('t Hart et al., 1990, p. 75). They "believe that the use of 'levels' in a phonetic analysis of intonation is an oversimplification. And even though it may be a commendable attempt at *phonological* data reduction, its application on the phonetic level runs counter to the phonetic facts of pitch-change production and perception" (ibid., p. 77).

Taylor (1994) criticized the IPO-approach as follows:

> "The Dutch system uses three rigidly defined levels, and therefore has problems dealing with any sort of downstep [see explanation on page 57, NB]. This strict three level distinction also poses problems with changing the pitch range or describing accent prominence [...].
>
> The phonetic, intermediate level is incapable of expressing all the necessary distinctions between downstepping and non-downstepping contours. [...] Thus the F0-intermediate and intermediate-F0 mapping are not the analysis and synthesis equivalents of each other [...].
>
> The fault in this case lies with "forcing" the F0 contour to be analyzed in terms of the three line declination system. If there is a large discrepancy between the behavior of real F0 contours and what the model proposes, then the model will run into severe difficulties. [...]

> The model will have difficulty analyzing any contour that is not within its own legal set" (Taylor, 1994, p. 27).

To what extent is the IPO approach useful for automatic detection of prosodic events? The approach is a possible technique to map the F0 level to more abstract intonational entities. However, the objections made by Taylor are crucial and before one attempts to implement the model in a fully automatic procedure a number of other questions remain to be solved, for instance how one can fit straight lines automatically with the same reliability as human labelers do it to the F0 curve. How can one further automatically process the stylized contour into a standardized contour?

Furthermore the stylisation with straight lines does not provide a level of abstraction as does a phonological model and can be criticized in this respect. The stylisation is more a sort of data reduction, whereas a phonological model enables one to structure acoustic observations and systematically explore patterns within those; that is, abstracting from the particular acoustic realization.

3.1.2 KIM - The Kiel Intonation Model

The KIM is an approach developed by Kohler and his coworkers (Kohler 1991, 1997) to model the intonation patterns occurring in German, although the basic prosodic categories should also be applicable to other languages. The description obviously focuses on speech synthesis and is also implemented in a TTS system. The model takes into account *microprosodic phenomena*[3] and also the fine detail of pitch movements within peaks and valleys expressed by the division of peak alignment into *early*, *mid*, and *late*. Kohler describes the model as follows:

> "KIM is integrated into a pragmatic, semantic and syntactic environment. The input into the model are symbolic strings in phonetic notation with additional pragmatic, semantic and syntactic markers. The pragmatic and semantic markers trigger, e.g., the pragmatically or semantically conditioned use of 'peak' and 'valley' types of sentence focus. Lexical stress position can largely be derived by rule, and syntactic structure rules mark deaccentuation and emphasis in word, phrase, clause and sentence construction. Phrasal accentuations are thus derived from the syntactic component preceding the prosodic model, and are given special symbolizations in the input strings to the model [...]" (Kohler, 1997, p. 190)

[3] Microprosodic phenomena are influences on the F0 contour resulting from segmental influences. Typical examples for a microprosodic influence are the short and sharp falls in the F0 contour after voiceless stops. Also the intrinsic F0 of vowels is subsumed under this label. See for instance Ladd (1996, p.284-285 footnote 7) and Laver (1994, p. 453 ff).

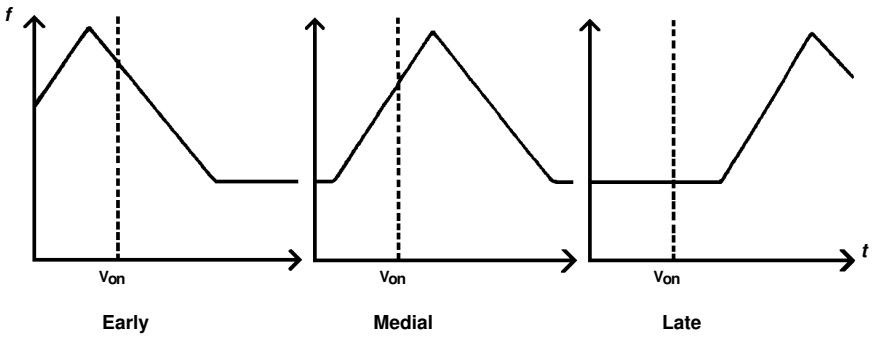


Figure 3.3: Illustration of early, mid and late peak alignment. The marker V_{on} indicates the onset of the stressed vowel (cited after Möbius, 1993, p. 48).

According to Kohler, KIM differs in two points from other intonation models. First, it integrates microprosodic phenomena and does not eliminate them. Second it incorporates pragmatic, semantic, syntactic, and meaning functions already at the building stage of the model. Kohler criticizes other approaches for postulating intonational units and thinking about their function afterwards.

The model separates two classes of rules: symbolic rules from subsequent parametric ones for generating acoustic output, which correspond to two levels of prosodic modeling:

1. the defining of phonology-controlled prosodic patterns by a small number of significant F0 points (macroprosody);

2. the output of continuous F0 contours influenced by articulation-related modifications (microprosody [...]) (Kohler, 1997, p. 190).

According to Kohler the alignment of the F0 peak with the segmental structure could be differentiated into three positions, each connected with a different meaning (see figure 3.3).

- "early: established fact; no room for discussion; final summing up of argument

- medial: new fact; open for discussion; starting a new argument

- late: emphasis on a new fact and contrast to what should exist or exists in the speaker's or hearer's idea" (Kohler, 1991, p. 125).

Möbius (1993, p. 49) criticizes Kohler with respect to this differentiation, because the functional meaning of the F0 peaks depends on the possibility of the speaker to produce the contours deliberate and the ability of the listener to identify the individual peak positions. "Astonishingly," continues Möbius, "Kohler himself seems to have serious doubts about this point, because during the production of the speech material that served as empirical basis for the development of microprosodic rules, only trained phoneticians were used" [my translations, NB]:

> "It had to be guaranteed that the global contours (early, medial, late peaks) stayed the same in all the sentence types. This precluded naive speakers as subjects because they are usually not able to keep a given utterance intonation constant throughout a whole experiment, which is, however, absolutely essential in the investigation of microprosody. Moreover, they even have difficulties with the realization of certain contours (e.g. early peaks)" (Kohler, 1991, p. 126 ff).

Because Kohler's approach is more concerned with how to get the right intonation for a given text, it is not well-suited for automatic recognition purposes. It is unclear how to get the parameters and input values from a given waveform. Kohler's model would need a full automatic speech recognition system that recognizes phonemes, syllables and words as to reconstruct the set of rules Kohler states. Since this is not yet solved satisfactory, it makes Kohler's approach less attractive for purposes of automatic prosody recognition.

3.1.3 Fujisaki's Model

The Fujisaki model (e.g. Fujisaki & Hirose 1982; Fujisaki 1983; Fujisaki 1997) was developed to handle the intonation patterns occurring in Japanese. However, it is intended to be applicable to any language, because it is based on the human production mechanism. The model grew out of the filter method first proposed by Öhman (1967). Fujisaki's basic assumption is that "F0-contours of words and sentences are generally characterized by a gradual declination from the onset towards the end of an utterance, superposed by local humps corresponding to word accent" (Fujisaki, 1981). Fujisaki uses a *phrase* and an *accent* component that are overlaid with each other to model intonation contours. Each of the phrase and accent components has a start time, an end time, and an amplitude. The model produces the F0 contour with a mathematical formula that is in the form of impulses, which in turn produces the phrase movements and step functions that calculate the accent shapes (see figure 3.4). Taylor states that Fujisaki showed "his model's operation on English, Estonian and Chinese intonation [...]" (Taylor, 1994, p. 40). Möbius (1993) has adapted the model for German intonation. However, Taylor's implementation of Fujisaki's model in a computer program for testing its usefulness for synthesizing English intonation patterns showed that the model worked well for neutral

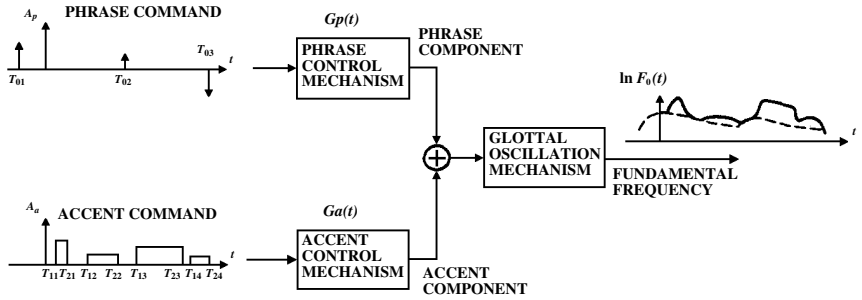


Figure 3.4: Sketch of Fujisaki's intonation model. A F0 contour is illustrated at the right-hand as the logarithm of the fundamental frequency, it can be approximated by the sum of the phrase and accent components. Phrase commands are series of discrete impulses and accent commands are a chain of rectangular impulses with varying height and duration. Both commands are approximated by the response of a second-order linear system to these commands.

declarative types of F0 contour, but, "it was difficult to be sure that the phrase shapes occurred in meaningful positions (i.e. near some kind of prosodic boundary)" (Taylor, 1994, p. 41). Ladd (1996, p. 30) also criticizes Fujisaki's model along this line, that is, to model certain intonation contours the phrase commands have to be put in places that make no sense linguistically. Möbius (1993) discusses several aspects of this in his thesis.

Hess et al. (1997) mentions Möbius et al. (1993), who developed an algorithm for automatic parameterization of F0 contours using the Fujisaki model. The algorithm was designed to reconstruct the phrase and accent parameters (Möbius, personal communication) and "which yielded good results for several categories of one-phrase and two-phrase-sentences" (Hess et al., 1997, p. 372). However, as there are neither subsequent notions of this algorithm nor any representative evaluation results, it remains unclear how the algorithms performance can be evaluated.

The linguistically unmotivated positioning of phrase accents is not the only problem with the Fujisaki model, it also runs into problems with downstepping contours and final falls. Liberman & Pierrehumbert (1984) mention this and demonstrate that the phrase *final lowering*[4] effect they discern cannot be captured even by using a negative impulse.

To use the Fujisaki model in an automatic recognition system would mean to reconstruct the underlying phrase and accent commands. Though Möbius et al. (1993) and others (e.g. Nakai et al. 1997; Mixdorff 2000) have developed approaches in this line the outputs are reconstructions of Fujisaki parameters but not of the underlying phonological form.

[4]Final lowering is a lowering and compression of the *pitch range* (the distance between the highest and the lowest point in the F0 contour of a speaker) in declaratives (cf. Liberman & Pierrehumbert 1984, Pierrehumbert & Hirschberg 1990, p. 278-279, and Ladd 1996, p. 77 ff).

3.1.4 Taylor's RFC-Model

Taylor presents a "phonetic model of intonation" that has three levels of description: a F0 level, an intermediate level and a phonological level. The F0 level is given by the continuous F0 values from a pitch tracker. The intermediate level introduces three basic elements of **rise**, **fall** and **connection** to model F0 contours. The phonological level uses **H** and **L** to describe high and low pitch accents, **C** to describe connection elements and **B** to describe the rises that occur at phrase boundaries.

Taylor wants to model both directions of mapping, F0 - phonology (analysis) and phonology - F0 (synthesis). Due to problems with the phonological level he only gives a fully specified grammar which links the intermediate and F0 levels.

Taylor's model grew out of Fujisaki's model. As a consequence of some problems with the modelling of specific intonation contours in the speech material analyzed, Taylor proposes a new accent component that has different rise and fall characteristics. Although the Fujisaki model is an *overlay model*,[5] the adaptation by Taylor results in a new model that uses a linear sequence of elements (rise, fall, connection). The connection element is intended to join the rise and fall elements and stems from Taylor's observation that "[...] most of the movement in the F0 contour occurred in the vicinity of its pitch accents. Except at the beginnings and ends of phrases, the F0 contour nearly always followed a straight line" (Taylor, 1994, p. 59). The number of possible accent shapes is also increased in Taylor's model. Taylor notes a similarity to the Dutch model, as his model also uses linear sequences of rises, falls and straight lines to model contours. However, his model is not constrained by the strict levels and declination lines of the Dutch model.

Taylor lays out his "New Phonetic Model" as follows:

- "F0 contours can be divided into a linear sequence of non-overlapping, contiguous *elements*.
- Each section is labeled with one of three fundamental elements: *rise*, *fall* or *connection*.
- The elements can occur in any order, with the exception that two connection elements cannot occur in sequence.
- Rise and fall elements are given by an equation. They can be scaled to any extent on the frequency or time axis.

[5] This term was introduced by Ladd (1988). "Overlay or superposition models treat the linguistic pitch contour as if it were some sort of complex function, which can be decomposed into simpler component functions" (Ladd, 1996, p. 24). Superposition models are contrasted to tone-sequence or autosegmental-metrical models where the course of F0 is determined by a sequence of phonologically distinctive tones. The general problem here is how the two approaches deal with local events and global trends in intonation contours. See also Ladd (1996, p. 24 ff) and Möbius (1993, p. 51 ff), for a discussion of this issue.

Type	Duration	Amplitude
rise	0.187	70
fall	0.187	-97
conn	0.175	0
rise	0.165	34
fall	0.100	-14
rise	0.171	57
fall	0.159	-93
conn	0.135	-7
silence	0.405	73
conn	0.105	0
fall	0.225	-76
conn	0.240	10
rise	0.175	43
fall	0.191	-57

Figure 3.5: Illustration of a RFC (**R**ise/**F**all/**C**onnection) description. From Taylor (1994, p. 64).

- Connection elements are straight lines of any gradient or duration.
- Fall elements are only used to represent falling parts of F0 contours which are associated with a pitch accent. All falling parts of F0 contours associated with pitch accents are represented by a fall element.
- Rise elements are used to represent rising parts of F0 contours which are associated with a pitch accent. All rising parts of F0 contours associated with pitch accents are represented by a rise element.
- Rise elements may also be used at the beginnings and ends of phrases where there is a sharply rising section of contour.
- Connection elements are used everywhere else; specifically to model parts of contour which do not have a pitch accent or a phrase boundary rise" (Taylor, 1994, p. 63).

The new intermediate description is a list of elements, each with a *type*, a *duration* and an *amplitude* (see figure 3.5).

An illustration of the output of the automatic labeler is given in figure 3.6.

Since the transformation of a continuous F0 contour into an abstract representation is an integrated part of Taylor's model, it should be one of the most fitting models for purposes of automatic prosody detection. However parts of the procedure chosen by Taylor are debateable.

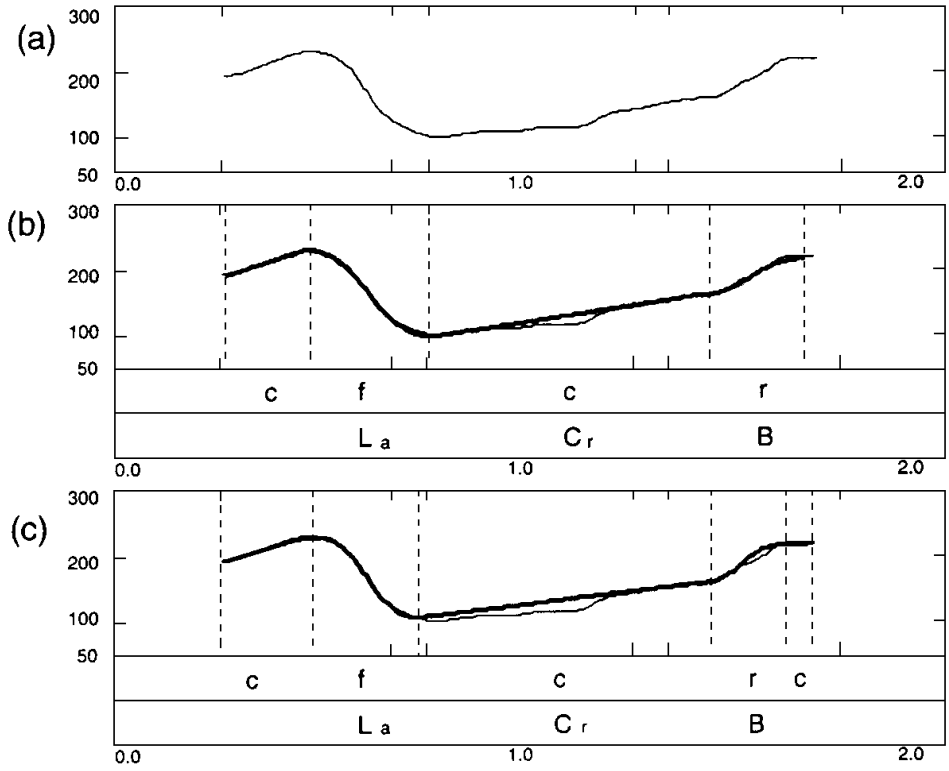


Figure 3.6: Example of a RFC-description of the utterance "Do you really need to win everything". In (a) the original F0 contour is illustrated, in (b) the manually labeled version and in (c) the automatically labeled one. The segments marked as "c", "f", and "r" correspond to the elements "connection", "fall", and "rise". The phonological elements "L_a", "C_r", and "B" correspond to a low pitch accent followed by a rising connection element and a final boundary. Here the automatic description and the hand labeled one are nearly similar (from Taylor (1994), p. 138).

The rise, fall, and connection elements are used to construct phonological elements which consist of four basic classes (H, L, C, B). H(igh) and L(ow) are used to classify pitch accents and may be subclassified by diacritics like 'd' for downstep in the case of H pitch accents. C stands for the connection element which Taylor characterizes as mostly irrelevant for the phonological description; though it is needed to reproduce the F0 contour accurately. However, as Taylor states, there are cases where the connection element has phonological function and these need to be labeled with the diacritic 'r(ising)'. The 'B' element is used to mark the boundary rises which could be continuation rises[6] or part of a combined accent construction such as a fall-rise. The diacritic 'i' for 'initial' marks boundary rises at starts of utterances. Taylor continues to formulate a finite-state grammar for his phonological elements. This grammar states that there must be at least one pitch accent in a phrase and that a phrase may start or end with a boundary element, and that connection elements can occur between any other elements.

Taylor states himself the question: Why do the elements have strict boundaries? He can offer no real explanation for it, but mentions that the curved starts and ends of the fall and rise elements are in fact transitions.

Taylor's model is based on decisions whether a specific part of the F0 track is categorized as one of the elements rise, fall, or connection. Problems arise by applying fixed thresholds for these elements. It is questionable whether these thresholds categorise too strict and subsequently label parts of the F0 track erroneously. Also smoothing of the original F0 track is used to eliminate microprosodic influences and heuristics are used to eliminate short deviations. At this point it remains unclear whether these processes introduce too strong manipulations of the original F0 track. Furthermore it can be criticized that only neighboring frames are compared and why there are no comparisons over larger domains.

3.1.5 Pierrehumbert's Model

Pierrehumbert proposed in her thesis (Pierrehumbert, 1980) a model of American English intonation that influenced a number of following approaches describing intonational phenomena. Her work is based on earlier studies from Goldsmith (1976), Leben (1976), and Liberman (1975). She describes an intonation contour as series of high (H) and low (L) tones. A set of diacritics distinguishes tones that are associated with accented syllables (marked with "*") from those associated with boundaries (marked with "%") and those between accents (marked with "–"). Her model forms the basis for the ToBI labeling instruction described in section 3.2.1. Her main aims were

> "to develop an abstract representation for English intonation which makes it possible to characterize what different patterns a given text

[6] See explanation on page 26.

> can have, and how the same pattern is implemented on texts with different stress patterns. The second aim is to investigate the rules which map these phonological representations into phonetic representations. These two aims go hand in hand, since we seek the simplest possible underlying representation by determining what properties of the surface representation can be explained by rules applying during the derivation instead of being marked in the underlying form" (Pierrehumbert, 1980, p. 10).

One of the innovations in Pierrehumbert's work was the possibility to make clear predictions as how to transfer an abstract phonological description of intonation into a concrete F0 contour. To produce F0 contours from an existing series of tones (which form pitch targets), interpolation rules are proposed. A declination baseline as well as a phonological downstep effect accounts for the downdrift observed in F0 contours. With regard to the other direction, that is the F0 – phonology mapping, Pierrehumbert states:

> "One consequence of our account of tonal implementation is that there is no level of systematic representation for intonation such as was suggested for segmental phonology in Chomsky & Halle (1968). This point can be made clear by considering the situation when the tone evaluation rules have gotton half way through implementing the tonal sequence for a phrase. To the right of the current window are the remaining unevaluated tones, still represented in the same form as in the underlying representation. To the left of the window is the F0 contour computed thus far (or a motor representation of it). The tonal sequence underlying this contour is entirely unaccessible; specifically, the types, locations, and phonetic values of tones are not accessed" (Pierrehumbert, 1980, p. 53-54).

In this statement Pierrehumbert denies that there is a way back from the continuous F0 values to the underlying tonal sequence. However, three years later Pierrehumbert presented a paper about "Automatic recognition of intonation patterns" (Pierrehumbert (1983), see a treatment of this paper in section 3.3.1) where she introduces an approach that analyzes F0 contours in terms of the theory laid out in her thesis. She states:

> "One aim of the project is to investigate the descriptive adequacy of this theory of English melody. A second motivation is to characterize cases where F0 may provide useful information about stress and phrasing. The third, and to my mind the most important, motivation depends on the observation that English intonation is in itself a small language, complete with a syntax and phonetics. [...] the F0

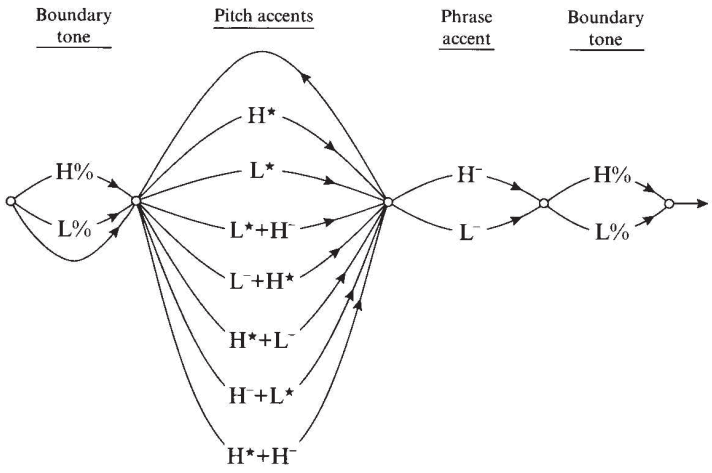


Figure 3.7: Finite-state grammar which "generates the set of well-formed tonal sequences for an intonation phrase." (Pierrehumbert, 1980, p. 29). "This grammar says that tunes are made up of one or more pitch accents, followed by an obligatory boundary tone. It implies two interrelated theoretical claims about the structure of tunes [...] First, the grammar implies that all possible combinations of pitch accents and edge tones are legal, [...] Second, it implies that there is no constituent structure to the contour, in particular no analogue to the 'head' and 'nucleus' of the traditional British analysis. Together, this mean that there is no difference between 'prenuclear' and 'nuclear' accents, except – trivially – their position: for Pierrehumbert the 'nuclear accent' is merely the last accent of the phrase" (Ladd, 1996, p. 81).

> contour, like other measurements of speech, is a continuously varying time function without overt segmentation. Its transcription is in terms of a sequence of discrete elements whose relation to the quantitative level of description is not transparent" (Pierrehumbert, 1983, p. 85).

However, since there was no later mentioning of a computer program that resulted from this concept it remains an open question whether this concept was successful or not.

According to Pierrehumbert (1980, p. 10-11) the phonological characterization of intonation consists of three components: (1) a grammar of allowable phrasal tunes (see explanation below figure 3.7), (2) a metrical representation of the text, and (3) rules for lining up the tune with the text. With regard to the metrical representation she refers to the metrical grid developed in Liberman (1975) and Liberman & Prince (1977) that describes which syllables are stressed and which are unstressed, and their relationship in strength among each other. At this point she notes that the strongest stress in the phrase the so called *nuclear stress*, will have an important role in the description of intonation. Figure 3.8 shows a description of an intonation contour in Pierrehumbert style notation.

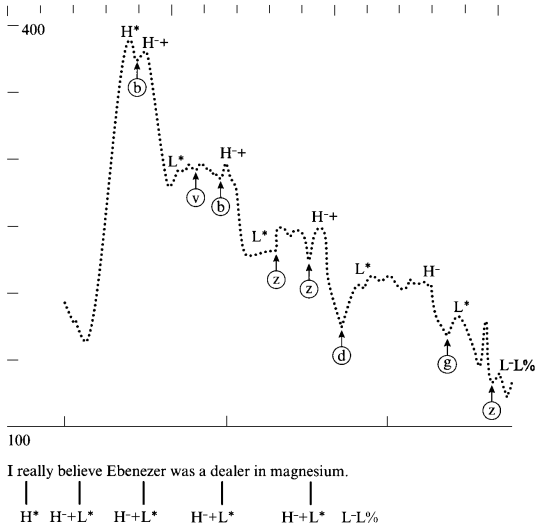


Figure 3.8: Typical description of an intonation contour in Pierrehumbert style notation (cited from Ladd, 1996, p. 86). The little circles including consonants indicate the positions of fricatives and stops that lead to a drop of the F0 contour (see section 5.1).

An important aspect in Pierrehumbert's theory is the claim that there are only two tone contrasts, H and L and not four as in earlier works (e.g. Trager & Smith 1951; Liberman 1975). In Hayes & Lahiri (1991) it is pointed out that this phonemicization resolves a serious difficulty faced by systems with more tones. "A system with, say four tones analyses certain pitch contours as distinct when they are arguably not. For example, a contour like 1 4 1 can have essentially the same meaning and usage as a 1 3 1 contour, differing only in degree of emphasis. A two-tone phonemicization represents both as LHL, allowing the paralinguistic factor of overall pitch range to determine the actual phonetic values." (ibid., p. 50).

Möbius (1993) criticizes Pierrehumbert's claim that intonation is determined only locally. He states that studies of utterances with parentheses have shown that the intonation contour is continued after the interruption nearly similar as it would be without the interruption (Möbius, p. 55 citing Kutik et al. 1983). It remains indeed unclear how the Pierrehumbert model would deal with such cases unless there is an explicit rule of how to continue an interrupted intonation contour at the same level as it was ended before the interruption. However, it is imaginable to continue the old contour or to start from the last F0 value which would imply to have some way of storage for this value. Taylor (1994) criticizes Pierrehumbert's model because the F0-phonology mapping "is very difficult to define in a formal manner". He argues that the interpolation rules are at fault and as a consequence "one must question the basis of the entire system" (ibid., p. 34). Since Taylor questions the central principle of the theory, namely that English intonation is a tone-based phenomenon, he argues against the Pierrehumbert system as basis for an automatic

prosody recognition program. Despite the status of the interpolation rules it is, however, indeed possible to construct a F0-phonology mapping procedure (as will be shown in section 4.4) under the assumption that the phonological system of pitch accents and boundary tones structures the incomming stream of features extracted from the course of F0 and other acoustic features. This mapping procedure does not refer to interpolation rules but incorporates rules that obey the phonological structure of tunes which include for instance restrictions in tone sequences.

Pierrehumbert's system provides both, an explicit description of how to transform a given sequence of tones into a concrete F0 contour, and a grammar of intonational tunes. That means the model encompasses the abstract phonological as well as the concrete F0 level of intonation. It has therefore explicit explanatory power regarding the phonological description of intonation but does not ignore the acoustic side of it. Particularly the ability to build abstractions from individual acoustic realizations towards a model of a few meaningful tunes relevant in a given language is one of the models advantages. However, since the model grew out of more theoretical considerations it is certainly more focused on the abstract level of intonation and there has been a lack of a successful F0 to phonology mapping.

When the model is able to adequately describe the intonation contours of a given language it should be possible to explore a way from the acoustic level to the phonological level and therefore to reveal the underlying intonation patterns by abstraction from the individual acoustic realization. The models explanatory power on the phonological level as well as on the F0 level combined with its concrete implementation into a intonation labeling instruction (see 3.2.1) qualifies the model as basis for an automatic prosody recognition program.

Another interesting aspect of the model are its predictions about the potential meaning of the pitch accents and boundary tones. Pierrehumbert & Hirschberg (1990) lay out a model of intonational meaning that provides further considerations regarding the relation between intonation structure and meaning. In this paper the authors describe tune meaning as compositional, that is composed of the combined interpretations of pitch accents, phrase accents and boundary tones. They propose "that a speaker (S) chooses a particular tune to convey a particular relationship between an utterance, currently perceived beliefs of a hearer or hearers (H), and anticipated contributions of subsequent utterances" (Pierrehumbert & Hirschberg, 1990, p. 271).

Hayes & Lahiri (1991) present an elegant application of the Pierrehumbert model for the phonological description of Bengali intonation. They argue that Bengali supports a typology of intonational tunes that includes only pitch accents and boundary tones. The phrase accent is reanalyzed as a boundary tone. Furthermore they show that Bengali intonation contours obey the Obligatory Contour Principle (OCP),[7] which forbids adjacent identical tones. By stipulating a phonological rule

[7]The Obligatory Contour Principle (OCP; Leben 1973; McCarthy 1986) forbids identical tones in sequence.

they convert underlying contours that violate the OCP to permissible surface forms. Additionally they show that Bengali phrasal stress assignment cannot be reduced exclusively to focus and other semantic factors as proposed by Bolinger (1972) but can be shown to have a default, phonologically assigned phrasal stress pattern. Hayes & Lahiri also mention "Gussenhoven's (1984) view that intonational tunes, just like segmental morphemes, may undergo phonological rules." (Hayes & Lahiri, 1991, p. 76-77).

Before the presented models are compared a short description of the autosegmental-metrical theory is presented since it provides important background information with regard to the Pierrehumbert model and as Ladd (1996, p. 111) points out: "The Pierrehumbert analysis of English is, in effect, one possible AM [autosegmental-metrical, NB] analysis among several."

3.1.6 Autosegmental-Metrical Theory[8]

Ladd (1996), states that the autosegmental-metrical theory started to develop in the late 1970s as an explicitly phonological approach to intonation based on the PhD theses by Liberman (1975), Bruce (1977) and Pierrehumbert (1980). This theory "adopts the phonological goal of being able to characterize contours adequately in terms of a string of categorically distinct elements, and the phonetic goal of providing a mapping from phonological elements to continuous acoustic parameters" (Ladd, 1996, p. 42). He also points out that the theory grew out of theoretical problems in phonology, which are partly addressed by the fundamental concepts of the autosegmental-metrical theory which Ladd explains by four basic points: (1) Linearity of tonal structure, (2) distinction between pitch accent and stress, (3) analysis of pitch accents in terms of level tones, (4) local sources for global trends. The *linearity of tonal structure* is represented by a string of local events (like pitch accents and boundary tones) associated with certain points in the segmental string. The pitch contour in between such events is phonologically unspecified and may be described with transitions. Although pitch accents serve as perceptual cues to stress or prominence there is a *distinction between pitch accent and stress*. Namely, pitch accents are "in the first instance intonational features, which are associated with certain syllables in accordance with various principles of prosodic organization" (Ladd, 1996, p. 42-43). The *analysis of pitch accents in terms of level tones* means, that level tones or pitch targets like H (high) or L (low) are the elementary units of pitch accents and edge tones in intonational languages. *Local sources for global trends* are explained by the operation of localized but iterated changes in scaling factors which result in overall trends in pitch contours.

In autosegmental phonology intonational features are represented on separate tiers where they may function partly autonomous. "Elements on the same tier are sequentially ordered, while elements on different tiers are unordered and related to

[8]This term is adopted from Ladd (1996, p. 42).

each other by means of association lines which establish patterns of alignment and overlap. Since associations between tones and tone-bearing units are not necessarily one-to-one, we may find other types of linking, [...]" (Clements & Hume, 1995, p. 247).

Some unresolved issues within the autosegmental-metrical theory according to Ladd (1996, p. 102 ff) are the following:

1. What is the status of pitch range?
2. What is a tone?
3. How are tones organized phonologically?

Regarding (1) - **what is the status of pitch range** - Ladd mentions that there are no fixed terms of reference for pitch like "high back unrounded vowel" or "bandwith of second formant 80 Hz" (Ladd, 1996, p. 252). He states that "the general problem of pitch range manifests itself within the AM [autosegmental-metrical, NB] approach as the specific problem of tone scaling" (Ladd, 1996, p. 272). As a solution of this problem he proposes a three-way classification of pitch range effects: (1) *Intrinsic* effects are those represented in the tonal string in terms of H and L tones, (2) *global extrinsic* effects - those based on the overall level and range of the speakers voice - are non-phonological, and should ideally be normalized out of phonetic data [...], (3) *extrinsic but linguistic* - such effects are not represented in the tonal string, but instead involve abstract relations between tones and between higher-level phonological constituents. Although this three-way distinction has interesting theoretical implications it is doubtful whether it could be incorporated without tremendous effort into a system of automatic detection of prosodic events. Normalizing speaker specific effects would imply to have some reference values already before any other estimations could be done. This certainly imposes severe restrictions onto an automatic detection system. Without neglecting the value of such approaches, it appears to be very difficult to successfully integrate it into a system that is able to process speech from different speakers and from different languages without any time-consuming preparation tasks.

With regard to (2) - **what is a tone** - Ladd mentions that "[...] pitch accents are composed of combinations of H and L tones. Yet fundamental questions remain about how tones are to be recognized: by what criteria do we decide that a given pitch accent consists of one or two tones? How do we determine that there is or is not a tone at any given point in a string?" Ladd (1996, p. 103). He continues with referring to "Bruce's [Bruce (1977, ch. 5), NB] original version of the AM approach, where tones are identified with turning points in the F0 contour. Local maxima correspond to H tone and local minima to L tone. [...] Also, Bruce (1977, ch. 5) explicitly discusses the possibility that tonal targets may be undershot in, for example, a H..L..H sequence where the targets are very close together. On

the whole, however, identifying tones with turning points puts severe limits on the range of possible phonological interpretations of a given contour" Ladd (1996, p. 103).

Ladd explicitly states that "This concrete conception of tones as turning points is implicitly abandoned by Pierrehumbert. In Pierrehumbert's approach, as we have seen, tones need not always correspond to turning points, and turning points need not always reflect the phonetic realization of a tone" (Ladd, 1996, p. 103). This has important implications for a model about automatic detection of prosodic events. Not only maxima or minima in the F0 contour may be possible pitch accent positions but other points as well. It is also important to know that undershot phenomena as described above may occur and that this cases are not entirely recoverable from the acoustic side. Though it is conceivable to introduce wellformedness constraints in order to force only specific contours being assigned. The latter opens another problem area, namely the one of overcorrection and acoustically unjustified tone assignments (cf. the discussion of this aspect below in section 4.4).

An important aspect for purposes of automatic detection is the "possibility that tones in the underlying phonological string may not be realized phonetically as distinct F0 targets" (Ladd, 1996, p. 104). This means that an automatic tone recognition program must have either a recovery mechanism that inserts a phonological pitch accent although there is no direct evidence in the course of F0 and RMS or has to accept the possibility that some tones may not be detected on basis of the F0 contour.

Regarding (3) - **how are tones organized phonologically** - Ladd discusses problems arising between the identification of tones and the analysis of their phonological organization, that is the phonetic - phonology relation. Specific theoretical assumptions could force a specific explanation of an intonational phenomena and could simultaneously produce problematic explanations. Ladd refers to work of Grice (1995) on Palermo Italian where she proposes a more complex structure of pitch accents that may consist of even three or four tones in a single pitch accent.

3.1.7 Comparison of Models

The models presented above show the diversity of approaches with regard to the modeling of intonational phenomena. Depending on their starting point the models emphasize different aspects. The IPO approach tries to reduce the observable F0 movements to only those, that are perceptually relevant. The perceptual relevance is provided by resynthesis of the stylized contour. It is, however, not obvious whether this procedure allows to cover all the intonational phenomena adequately because the judgment of perceptual equality is subjectively based. Moreover, the condition for intonation contours to fall into the three declination lines may cause problems with contours that deviate extremely from this predefined lines.

Fujisaki's model has an appealing aspect, namely its basement in the human articulation process. It was also used in a number of speech synthesis programs. Despite this facts, the model has problems with the placement of its accent and phrase commands in linguistically meaningful positions. It has been shown that the method was not able to model some typical intonation contours occurring in English or was not able to produce certain contours without placing some commands at linguistically unmotivated positions.

The Kiel model of intonation is attractive because it incorporates different sources of intonation determiners (i.e. pragmatic, semantic and syntactic aspects). It is, however, more focused on the synthesis aspect than on the analysis side. It remains unclear how one could integrate all the factors in an automatic analysis tool. Further the relevance of the peak alignment positions are questionable.

Taylor's RFC-model offers an explicitly formal approach for intonational modeling and includes a F0-phonology mapping process. Therefore it comes closest to the requirements for an adequate model for the automatic analysis of prosody. The method itself takes only the F0 track into account and does not cover the interaction of the parameters F0, duration and intensity. Therefore it is very sensitive to the course of F0 and consequently can be misguided by erratic parts in it. Especially the procedures chosen to minimize the segmental effects upon the F0 contour could (a) smooth the original contour too much and therefore delete possible accents, or (b) smooth too less and therefore allow some faulty tone assignments, or (c) could produce new F0 movements not originally present (for instance at transitions between voiced and unvoiced parts) and subsequently introduce faulty tone assignments. It is doubtful whether the sole inspection of F0 provides sufficient acoustic features for a reliable detection of abstract phonological entities.

Pierrehumbert's model originates from the phonological side but gives an explicit description of the transformation from the abstract to the F0 level of representation. The model has its advantages in its descriptive force, and the ability to abstract away from peculiar acoustic differences towards the more general structure of intonational phenomena in a given language. The model has been successfully applied to the description of different languages and is used by a large number of researchers of the linguistic community.

The decision for using the Pierrehumbert model of intonation as basis for the automatic detection of prosodic events is based on the following considerations:

- the models descriptive force
- can be used to make predictions about possible intonation contours (rather than only describing them)
- its widespread usage around the world
- its potential for an universal description system for intonational phenomena

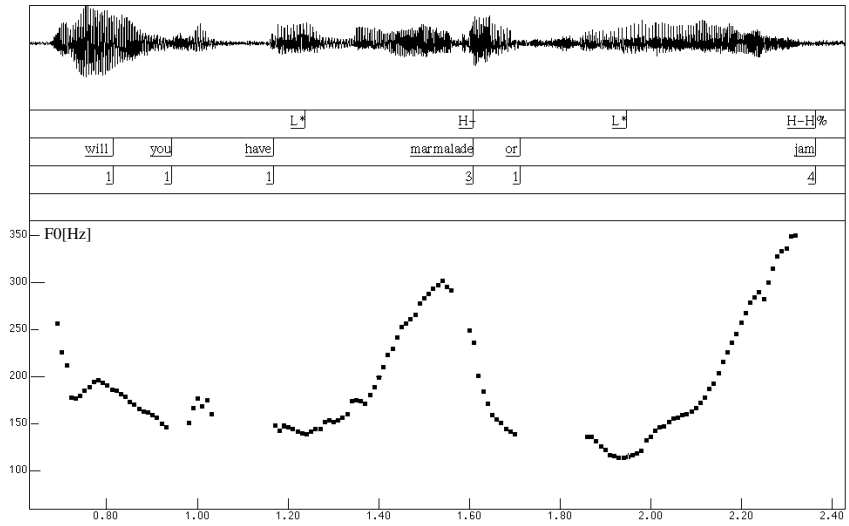


Figure 3.9: Example for an intonation contour labeled with the ToBI conventions. The first label tier represents the tone-tier, the second one the word boundaries, the third one the break indices and the fourth one a miscellaneous tier. This is one of the examples in the publicly available ToBI material (Example: "jam1"; Beckman & Ayers 1997, p. 1).

- the authors own experiences with the model while labeling speech data

This decision, however, does not prevent other models to be taken as phonological back end for the method presented in this thesis. The next section will deal with the implementation of the Pierrehumbert type model in a labeling instruction and contrasts it with another labeling method.

3.2 Labeling Methods

3.2.1 ToBI

ToBI (= **To**ne and **B**reak **I**ndices) is a labeling instruction for "transcribing the intonation patterns and other aspects of the prosody of English utterances" (Beckman & Ayers, 1997, p. 1). The ToBI labeling conventions are based on the intonation model of Pierrehumbert (Pierrehumbert 1980; see 3.1.5). The model uses pitch accents, phrase accents and boundary tones as determiners for the intonation contour of a given utterance. Pitch accents are either high (H) or low (L) tones or combinations of both, but only one of them bearing the star (e.g., H*+L, L*+H) indicating that it is associated with the lexically stressed syllable. Phrase accents are marked with the diacritic "-" (e.g., H- or L-) and said to be responsible for the movement of the F0 contour between the nuclear accent and the boundary tone. Boundary

tones are either H% or L% when occurring at the end of an intonation phrase or %H, %L, or missing when occurring at the beginning.

The phrase accents were proposed in the original Pierrehumbert thesis (Pierrehumbert, 1980), but later (Beckman & Pierrehumbert, 1986) changed with the introduction of a second, smaller level of phrasing (than the intonation phrase), namely the *intermediate phrase*. However, the phrase accents were reintroduced in the ToBI labeling guide (Beckman & Ayers, 1997).[9]

The ToBI labeling conventions use 4 different label files:

1. A **tone label file**, including the pitch accents, phrase accents and boundary tones.

2. A **label file for the words** and word boundaries, spoken in the labeled waveform.

3. A **break-index label file**, using 4 different levels of strength of pause between words, intermediate phrases or intonation phrases, ranging from 1 for the weakest break up to 4 for the strongest break usually indicating the end of an intonation phrase.

4. A **miscellaneous label file**, which includes events like coughs, disfluencies, interruptions, faults in the recording, etc.

Figure 3.9 shows an example of a ToBI-transcription taken from the ToBI labeling material.

What are the conventions for placing labels? Phrase accent and boundary tone labels should be placed at the end of an intermediate or intonation phrase. Pitch accents should be placed within the accented syllable, preferably within the syllable's vowel (see section 1.5 "What lines up with what?" in Beckman & Ayers (1997) and also the definition of tones in GToBI in section 3.2.3).

Each individual tone proposed in the underlying inventory is described with concrete characteristics regarding its acoustic features and its placement in time. According to Beckman & Ayers (1997) a crucial distinction is the one between L*+H and L+H*. In the L*+H accent ("scooped accent") the low tone is associated with the stressed syllable and the high tone occurs much later (one or more syllables). Opposed to that, the L+H* accent ("rising peak accent") has the low tone much earlier and the stressed syllable is associated with the high tone. Though descriptions like the latter appear to be plausible, they are sometimes not unproblematic when labeling new data.

[9] See also the discussion of the ToBI labeling conventions and the status of the phrase accent in Ladd (1996, p. 94 f).

Labeling acoustic data according to the ToBI-conventions is time-consuming and although an experienced labeler develops more confidence it is in some cases hard to decide whether a given F0 movement falls in the one or the other category. A typical problem case is the differentiation between H* and L+H*. Although there are investigations of inter-labeler consistency that have shown good agreements (see e.g. Grice et al. 1996) it was also expressed that "manually labeled speech corpora may not be sufficiently consistent for successful training or modeling for recognition or TTS systems" (Syrdal & McGory, 2000, p. 4). At this point it suggests itself to base the labeling on more objective criteria and subsequently use an automatic procedure for it. When the algorithm relies on objective criteria only and is not biased by linguistic intuition it should therefore be able to separate tones with robust acoustic cues from tones with less robust cues.

In conclusion, the ToBI labeling conventions are well described in the ToBI label guide (cf. Beckman & Ayers 1997) and language specific adaptations of it have been made for several languages (e.g. Bengali, Dutch, German, Greek, Japanese, Korean; cf. section 3.1). One of the advantages (but simultaneously one of the problematic aspects) of the ToBI labeling system is certainly its abstraction from the acoustic detail and its concentration on the phonological level of prosodic transcription. Using the ToBI labeling conventions for labeling large acoustic speech corpora is not unproblematic as a result of the fairly complex interaction of theoretical assumptions underlying the model, detailed labeling instructions with respect to the type and placement of tone labels and its application to highly variable acoustic data. Labelings from different human labelers have been criticized to be not consistent enough for their usage in automatic speech processing techniques (Syrdal & McGory, 2000, p. 238). Despite its problematic aspects, the ToBI system has been shown to account for the description of a number of intonational phonologies around the world and is therefore certainly qualified as the backend of an automatic prosody detection system.

3.2.2 INTSINT

Compared to the ToBI labeling conventions that were first published in 1992 the INTSINT (= **IN**ternational **T**ranscription **S**ystem for **INT**onation) transcription system is a more recent development. INTSINT was developed to be an "equivalent of a narrow phonetic transcription and can consequently be used for gathering data on languages which have not already been described" (Hirst & Di Christo, 1998, p. 14). The authors state that the original motivation for INTSINT was to develop a system that could be used for transcribing both English and French. INTSINT uses a set of transcription symbols to mark static points (the authors refer to "turning points" of Gårding 1977) which are assumed to be the most adequate phonetic representation (but see the critique of Ladd regarding turning points in section 3.1.6 and in Ladd 1996, p. 103 ff).

		Positive	***Neutral***	***Negative***
ABSOLUTE		T [$\Uparrow$]	M [$\Rightarrow$]	B [$\Downarrow$]
RELATIVE	*Non-iterative*	H [$\uparrow$]	S [$\rightarrow$]	L [$\downarrow$]
	Iterative	U [$<$]	$\bullet$	D [$>$]

Table 3.1: The inventory of description elements in the INTSINT system. "The letters stand for Top, Mid, Bottom, Higher, Same, Lower, Upstepped and Downstepped respectively." (Campione et al., 2000, p. 190).

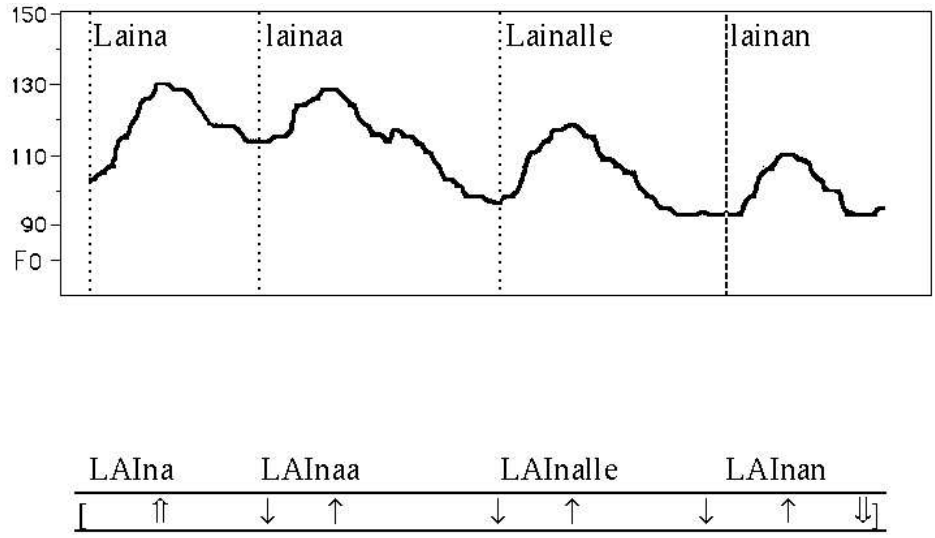


Figure 3.10: Illustration of the transcription of the Finnish sentence *Laina lainaa Lainalle lainen* (Laina lends Laina a loan) with the INTSINT labeling method. "The relative scaling of the points within an Intonation Unit need not be specifically marked since it is assumed that the most important factor is the height of each successive point relative to the previous point." (Hirst & Di Christo, 1998, p. 16)

INTSINT intends to provide a formal description of a F0 curve and is therefore unlike ToBI, which describes events of a linguistic nature. The pitch symbols in INTSINT represent a pitch **point** or **target**. Each of these targets is specified in one of two ways: "[...] either as an absolute tone, defined globally with respect to the speakers pitch-range, or as a relative tone, defined locally with respect to the immediately neighboring target-points. Relative tones can further be subdivided into iterative and non-iterative categories where it is assumed that iterative tones can be followed by the same tone whereas non-iterative tones cannot" (Campione et al., 2000, p. 189) (see the inventory of INTSINT labels in table 3.1 and an example of an INTSINT transcription of a Finish sentence in figure 3.10). The analysis therefore evolves from speech signals without the application of specific language models. INTSINT may also be used for synthesis purposes, where it needs some language knowledge to generate F0 contours from the grammar-generated target points.

Hirst and Di Christo's approach seems to offer a possibility of unifying intonational transcriptions across languages. However, there is no obvious improvement with this transcription system with respect to the basic structure of the phonological system. As was pointed out in section 3.1.6 it is insufficient to associate pitch accents only with turning points in the F0 contour. INTSINT might provide a closer phonetic description of F0 movements, but it is unclear whether it is able to posit abstractions from phonetic detail in the same way as the autosegmental-metrical approaches are able to.

Because first experiments with the approach presented in this thesis were carried out on German speech material, the next section will present an adaptation of the ToBI model to German.

3.2.3 GToBI

GToBI stands for German ToBI and was developed in cooperation by Martine Grice, Matthias Reyelt, Ralf Benzmüller, Anton Batliner and Jörg Mayer (Grice et al., 1996). The labeling system has recently been modified in order to make it phonetically more transparent and to integrate recent advances in intonational phonology (Baumann et al., 2001). Grice et al. (to appear) presents a review of accounts for German intonation and a motivation for the GToBI model. In the following the individual pitch accents and boundary tones are listed with a detailed description from the GToBI labeling guide. GToBI has two monotonal pitch accents (Grice & Benzmüller, 1997):

H* The accent forms the peak of a more flat increase. There is no low or high target before the accented syllable. It is the default accent that is used when there are no clear indications for other accents. (see two examples in figure 4.1 in the following chapter); the so called 'peak

accent' is characterized by an upward movement of the pitch. The accented syllable sounds high. The upward movement is not as high as in the L+H* case.

L* 'Low accent' an apparent tone target on the accented syllable low in the speaker's range, often corresponding to a dip in F0; characterized by a downward movement of the pitch or a low register. Usually the minimum of the pitch movement lies approximately in the middle of the vowel within the accented syllable.

There are four bitonal pitch accents proposed by GToBI:

L*+H A minimum is reached on the accented syllable. There will be an increase late in the accented syllable that ends in the following syllable (sometimes later). The accented syllable sounds low.

L+H* This pitch accent is characterized with a steep increase on the accented syllable. The endpoint of the increase is late in the accented syllable sometimes afterwards. Important is that the accented syllable sounds high. Important as well is that the syllable before the accent reaches a low target (see an example in figure 4.2 in the next chapter).

H+L* The accented syllable is low. The preceding one is high. The impression of a large pitch jump towards a low level appears.

H+!H* Here a fall on the accented syllable from a high to a mid level of the speakers range occurs. The syllable preceding is higher. If H+!H* is followed by a L- boundary tone then the pitch decreases further. If it is followed by a H- pitch then it stays on the level of the !H*.

The above mentioned pitch accents may be labeled with an additional diacritic signaling "upstep" or "downstep" of the tones. Downstep is a lowering of an assumed topline of pitch range which shifts the F0 of an H accent downwards and is marked by placing a "!" symbol in front of the affected H tone. Upstep is used to indicate a step up within a sequence of pitch accents and also describes a step up to a boundary tone. Each accent of such a sequence is marked with a "^" symbol.

There are altogether four intonation phrase boundary tones and two intermediate phrase boundary tones proposed by GToBI. Boundary tones have to appear at the end of a phrase and may also appear at the beginning of phrases although the latter is usually unmarked. Two types of phrases are separated:

- intermediate phrase = ip = small boundary,
- intonation phrase = IP = large boundary.

An intonation phrase consists of minimally one intermediate phrase and includes therefore always two boundary tones.

Boundary tones determine the course of intonation between the last accent (nuclear accent) and the end of a phrase. Boundary tones are labeled at the end of the last word in the phrase. This does often not correspond to the pitch contour and is just a convention. The two phrase accents are:

L- Phrase accent occurs at the end of an intermediate phrase. L- describes two parts of the pitch movement after the accent: (i) after H*, L+H*, H+!H*, L*+H a decrease in to the lower range of the speakers voice will be immediately following the accented syllable, (ii) the continued decrease of the contour towards the low baseline, that mostly ends on a following stressed (but not accented) syllable.

H- Phrase accent occurs at the end of an intermediate phrase. H- after H*, L+H*, H+!H* is realized with a continuation intonation. The pitch stays on the level of the H accent tone (or !H tone) or slightly above; it might also decrease or increase slightly. The pitch movement does not follow the high register line, an assumed top line of the speakers pitch range. H- after L* results in an increase to a middlehigh level that is reached either at one of the following stressed syllables or at the end of the accented word. Afterwards there will be a plateau reaching to the end of the intermediate phrase.

Phrase accents which function as edge tones but may also associate with stressed syllables or other tone-bearing units can now be marked with a separate L(*) or H(*) label, to indicate the secondary nature of the postnuclear prominence (cf. Grice et al., to appear). The four final and the one initial boundary tone are:

L-% Marks a larger boundary than L-. The decrease in pitch is most often lower than in L- alone. After the falling movement that appears only after H accents the contour is flat along the lower register line. On the last syllable pitch is decreasing even more below the register line (so called “final lowering”, see page 39). This boundary tone was formerly “L-L%”. It was reduced to the current version of GToBI in order to make it phonetically more transparent. When intonational phrase (IP) and intermediate phrase (ip) boundary tones would represent the same pitch level, only one tone is transcribed

L-H% Following H accent tones the L-H% is a fall-rise movement. After the minimum of the fall is reached the contour stays on a low level. On the last syllable (or shortly before) there is an increase until approximately to the mid of the speakers range, by which the high register line is mostly exceeded. After L- accent tones the L-H% is a concave rising movement that starts only on the last syllable.

Pitch Accents	Boundary tones
H*	L-
L+H*	L-%
H+!H*	L-H%
L*	H-
H+L*	H-L%
L*+H	H-%

Table 3.2: Inventory of pitch accents and boundary tones in GToBI.

H-% The pitch movement in H-% is not separable from H-. After the H- the level of the following boundary tones is increased so much that no significant decrease appears. H-% does not mark a falling movement. Formerly "H-L%".

H-H% After a H accent tone the H-H% will be first realized with a continuous intonation, afterwards it increases on the last syllable. The increase on the last syllable reaches the highest regions of the speakers range. H% after a H- tone is higher than all the other H- tones. Following a L* or H+L* the H-H% is realized with an increase high into the speakers range. The increase is convex, that is after the accented syllable there is a middlehigh increase. After the plateau an increase occurs on the last syllable up to the highest levels of the speakers range.

%H Initial high boundary tone.

The "H" in each boundary tone may also be labeled with upstep or downstep diacritics.

The inventory of pitch accents and boundary tones in GToBI is summarized in table 3.2. Boundary tones are realized on the end of intonation phrases and determine the phrasing. However, boundary tones can also occur at the beginning of a phrase. As already mentioned in 3.2.1 the status of the phrase accent is not without problems. In the theory of Hayes & Lahiri (1991) there is a notational addition to the system used in Pierrehumbert, namely

> "The notion of the intermediate phrase has been adopted by numerous authors, though the term 'intermediate' poses a problem if for no other reason than that it makes it difficult to come up with unambiguous abbreviations for intermediate phrase and intonation phrase. Hayes and Lahiri (1991) suggest that the intermediate phrase is equivalent to the 'phonological phrase' of Nespor & Vogel (1986) and others. They also propose a potentially useful notational device, based

> on Beckman and Pierrehumbert's idea that the phrase tone is the edge tone for the phonological phrase and the boundary tone that for the intonational phrase; in place of Pierrehumbert's T for phrase tone and T% for boundary tone, they write T_p and T_i respectively. A similar notation is used by Grice (1995) who writes T_b for intermediate phrase edge tones (phrase tones) and T_B for intonation phrase edge tones (boundary tones). However, neither notification has so far been widely adopted, and as long as there remains doubt about the proper analysis of falling nuclear accents (and more generally about the status of the phrase tone), the notational question seems likely to remain unsettled [...]" (Ladd 1996, p. 93-94).

In her thesis, Féry (1993) argues against the phrase accent in describing German intonation patterns, by claiming that "the functions it fulfills in English can be represented by the trail tone of a bitonal tone and the phrasing, which exists independently of the tonal structure anyway" (Féry, 1993, p. 72) Therefore all nuclear accents in Féry's approach are bitonal, whereas prenuclear tones can be monotonal.

Despite the controversy about the phrase accents the GToBI model is one possibility of describing German intonation within the autosegmental-metrical framework. It has the advantages that the theoretically stated pitch accents and boundary tones are represented by concrete examples in accompanying training material (Grice & Benzmüller, 1997) and is surface-oriented which is important for the purposes dealt with in the present thesis. Since an approach about automatic detection of prosodic events has to have explicit reference to concrete examples of these prosodic events in speech data, the GToBI model is well suited as basis for this purposes.

After having laid out the theoretical background of intonation analysis the next section will review some of the existing approaches concerning automatic prosody recognition.

3.3 Existing approaches about automatic recognition of prosodic events

This section will give an overview of existing approaches about automatic recognition of prosodic cues and will discuss them. The following approaches will be presented in detail: Pierrehumbert (1983); Wightman & Ostendorf (1994); Taylor (1994) (RFC-model); Rapp (1996); Ostendorf & Ross (1997); Hirst et al. (2000) (MOMEL), and Wightman et al. (2000) (ToBI Lite).

3.3.1 Pierrehumbert

In Pierrehumbert (1983) an approach for "automatic recognition of intonation patterns" is presented that is greatly influenced by work of Marr (1982) and his collab-

orators on vision. “The schematization of the F0 contour has a family resemblance to their primal sketch, and I follow their suggestion that analysis of derivatives is a useful step in making such a schematization” (Pierrehumbert, 1983, p. 89).

The taskflow in Pierrehumbert’s approach is as follows: input to the system is a F0 contour that is processed in two respects, first continuity constraints are applied in order to prevent F0 values resulting from pitch doubling or halving errors to be taken as serious. Such values could even be deleted when the mechanism fails to bring it into line. Second the microprosodic effects on the F0 contour were tried to be eliminated by removing F0 values in the immediate vicinity of obstruents. Then the retained portions of the F0 contour were connected by linear interpolation. Then the “connected contour is smoothed by convolution with a Gaussian in order to permit the analysis of the derivatives” (ibid, p. 89). Afterwards events in the contour are detected by the analysis of the first and second derivatives. “The events of ultimate interest are maxima, minima, plateaus and points of inflection. Roughly speaking peaks correspond to H tones, some valleys are L tones, and points of inflection can arise through downstep, upstep or a disparity of prominence between adjacent H accents” (ibid, p. 89). Finally the tone association is provided by a “topdown nondeterministic finite state parser, assisted by a set of verification rules” (ibid, p. 89).

Interestingly one of the main problems arising in Pierrehumbert’s approach results from the smoothing introduced to suppress segmental effects. First of all it results in poor time alignment and secondly “curves that are too smooth may still be insufficiently smooth to parse” (ibid., p. 90). She also states: “Thus, I view the separation of segmental and prosodic effects on F0 as an open problem. Adding verification rules for segmental effects appears to be the most promising course” (ibid., p. 90) It is interesting to note that exactly the last mentioned aspect, that is the separation of segmental and prosodic effects, will play an important role in the approach presented later in this thesis (see section 5.1).

However, since there were no detailed performance results given nor any later mentioning of this approach it remains unclear whether it was ever successfully integrated into a working system.

3.3.2 Wightman and Ostendorf

Wightman & Ostendorf (1994) present an algorithm for labeling prosodic patterns in speech which uses HMM techniques to process the waveform for an automatic segmentation before they apply the intonational analysis procedure. By applying the automatic recognition of words and subsequently of word boundaries they gain the advantage to use phrase final lengthening and other durational cues during their intonation pattern recognition procedure.

They use four intonation markers: “ “P” for prominent syllables, “s” for unmarked syllables, “BT” for a syllable marked with an intonational phrase [...] and “P-BT”

for syllables marked with both a prominence and a boundary tone" (ibid, p. 471). Wightman & Ostendorf list a number of acoustic cues of prosodic events that have been described in the literature. With regard to acoustic correlates of prosodic phrase boundaries they state the following: "boundary foot lengthening, preboundary lengthening, pauses, breaths, boundary tones, and speaking rate changes" (ibid, p. 472). Of these, they consider all but foot lengthening, since it is a consequence of preboundary lengthening. As other cues of phrase boundary marking they note unfilled pauses and breaths. However, they also state that pauses are not completely reliable indicators of major phrase breaks since many intonational phrase boundaries are not followed by pauses and in spontaneous speech, pauses could be a result of hesitations and are therefore not a prosodic break. A further cue which is always present at a major boundary is the boundary tone which they define as "distinctive f0 features that signal major phrase boundaries." (ibid, p. 472). Finally they mention speaking rate changes as cue for signaling phrase boundaries. They propose a normalized duration measurement developed in Wightman et al. (1992) that averages over several syllables for the estimation of speaking rate.

As cues for phrasal prominence marking they mention: "[...] duration lengthening, pitch accents, and increased energy [...]. The relative importance of these cues and their interaction is not well understood [...]" (Wightman & Ostendorf, 1994, p. 472).

The reduced set of classification labels (syllables may be labeled as: s = unmarked, P = prominent, BT = boundary tone, and P-BT = prominent + boundary tone) as compared to the full set of ToBI labels (5 pitch accents, 2 phrase accents and 2 boundary tones, ignoring downstepped variants) is certainly more easier to cover and also a reason for the detection accuracy they get. The authors evaluate their algorithm on two corpora consisting of professionally read speech. The first corpus consists out of 280 sentences or 2140 words and includes 4 different speakers. The second corpus is a collection of radio news stories from one female speaker and contains 457 sentences or 8568 words. Both corpora were manually labeled with the prosodic labels and automatically segmented with phoneme labels. The latter was provided by a recognizer constrained by the known word sequence. Wightman and Ostendorf report an overall accuracy for the first corpus of 79%. "Prominences (combining P and P-BT) are correctly detected at a rate of 83% and falsely detected at a rate of 14%. Boundary tones (combining BT and P-BT) are correctly detected at a rate of 77%, but have only a 3% false detection rate" (ibid, p. 477). The second corpus showed similar accuracy results.

The approach presented by Wightman & Ostendorf needs the knowledge about the sequence of phonemes, syllables and words in the speech analyzed. This is provided by forced alignment of the known word sequence, but is, although fairly advanced and reliable, not flawless. Therefore faulty segmentations are transferred into the intonational analysis and produce accumulated errors. By introducing the automatic segmentation before the intonational analysis one gains more restricted search domains and access to durational features in the segmental domain but on

the other hand has to deal with faulty segmentations. Furthermore only controlled recording conditions and a restricted number of speaking styles was used in the evaluation corpora. As already mentioned above, the limited set of labels especially with regard to syllabic prominence (either unmarked or prominent) certainly reduces the complexity enormously. It has also been stated later by Ostendorf & Ross (1997, p. 304) that the error rates increased by 20-50% when labeling speaker-independent spontaneous speech with this approach.

3.3.3 Taylor (RFC-Model)

Taylor's model was already discussed in section 3.1.4, but in the following the model's specific parts regarding the automatic analysis of F0 contours are focused on. Taylor only processes the F0 contour to extract automatically a phonological description of an utterance's intonational tune. He uses three elements to classify F0 movements: rise, fall, and connection and calls the analysis module consequently "Automatic RFC Analysis System" (Taylor, 1994, p. 90). This system includes three main modules: (1) **contour preparation**, which includes extracting and processing the F0 contour in order to reduce segmental influences; (2) **broad classification**, which labels a given utterance into the three sections rise, fall, and connection of which the boundaries are only loosely defined; and finally (3) **optimal matching**, which determines the exact boundaries between the sections. It is remarkable that Taylor uses F0 contours extracted from laryngograph recordings. Usually these are only available from recordings under laboratory conditions and F0 is much more easily detectable (as the author mentions himself on page 54 of his thesis).[10] Therefore the probability of erroneously extracted F0 values is largely reduced which in turn makes the usage of the algorithm on basis of F0 values extracted from waveforms not impossible but as the author mentions later, performance is about twice as bad as on laryngograph F0 contours (cf. Taylor 1994, p. 115).

Though laryngograph F0 contours were used a considerable amount of work is spent for the contour preparation. Since there is no detection of segmental content but segmental effects on the F0 contour are present the question came up how to remove these effects and still leave the meaningful parts in the contour? Taylor decides to apply a 15-point median smoothing, which replaces each F0 value with the median of itself and 14 neighboring values, seven on each side. This procedure is used to smooth both pitch perturbations as well as segmental effects in one step. Afterwards the unvoiced regions are linearly interpolated and once again smoothed by a 7-point median smoothing in order to remove singularities at the voiced to unvoiced and vice versa joins. Though this procedure showed to be essential for Taylor's approach it is debatable whether it introduces too much

[10] However, laryngograph recordings are also not without problems, e.g. subjects should avoid fast movements, external electromagnetic fields might interfere with the laryngograph signal, also vocal fold movements without vocal fold contact are not measured (cf. Reetz 1996, p. 86 ff).

smoothing and subsequently levels out important F0 movements or as another consequence produces misleading F0 movements that are later erroneously labeled as meaningful.

After the contour preparation took place the broad classification module was applied. F0 values were originally be extracted every 5 ms but re-sampled so that a value was present every 50 ms. The latter is justified by the circumstance that intonational information varies more slowly than the original 5 ms steps and in order to keep the size of information presented to the system as small as possible. The broad classification module categorizes the single F0 values into rises, falls, and connections by comparing them to the previous F0 value. When the ratio is higher than a certain threshold the frame is labeled as a rise, when it is lower it is marked as fall. The remaining cases are labled as connection. Originally these thresholds were defined manually, but later optimized. When there was a series of frames belonging to the same class they were grouped together to form a labled 'section'. Possible errors of this method are: labeling a rising connection element as rise when the rise threshold was to low (the author calls this an 'insertion'); when the rise threshold was too high, then legitimate rises could be labeled as connection ('deletion'). Both cases could also appear in case of the fall elements. The optimal threshold values would have the number of these errors minimized (cf. Taylor 1994, p. 95). With respect to this method it remains questionable whether it applies too strict classifications which subsequently are not sufficient to categorise F0 movements into meaningful and non-meaningful parts.

The output of the broad classification module was used as input for the optimal matching module "which was designed to find the precise boundaries between the sections" (Taylor, 1994, p. 97). The optimal matching module went back to the original 5 ms step size of the F0 contour in order to find the precise boundaries beetwen the broadly classified sections. To find the precise boundaries candidate rise and fall shapes of different duration and amplitude were synthesized and compared with the original contour. "The shape that fitted the contour best was kept, and its start and stop times were taken as the precise boundaries of the element" (ibid., p. 97). The matching procedure, although in itself an interesting approach, is fairly complex since about "400 monomial[11] shapes had to be calculated for every element, and each of these shapes had to have the distance between itself and the F0 contour measured" (ibid., Appendix D). It is questionable whether such a complex procedure is justified for such a task.

Taylor continues to emphasize that the segmental influence upon the F0 contour is the reason for incorrectly recognizing pitch accents in his approach. Though this is certainly a major challenge in an automatic prosody analyzer it remains still an open question whether the knowledge of segmental content would significantly improve his results since it is unclear whether (a) the three-way classification of

[11] "A monomial is a function of the form $y = ax^n$, i.e. a function with only one x term" (Taylor, 1994, p. 58).

F0 movements into fall, rise, and connection elements is sufficient for recognizing pitch accents, especially without the knowledge about the simultaneous energy contour, and whether (b) the optimal matching procedure proposed is too complex and still insufficient for a model of F0-phonology mapping.

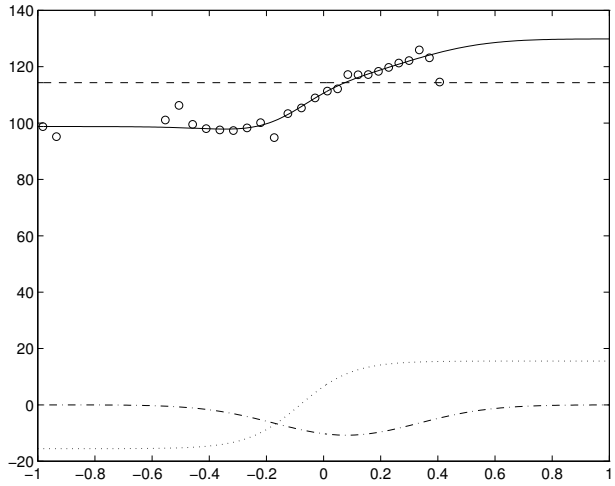
Though the analysis of Taylor's approach has revealed a number of problematic aspects in it, it shows clearly the basic problems of an automatic intonation analyzer. One of the most important issues is the separation of segmental and F0 tracking specific influences on the F0 values from the potentially meaningful parts. Another aspect involves the procedure of how to map from the (quasi-)continuous F0 values to the discrete phonological elements in order to cover the full range of possible contours in the speakers set.

3.3.4 Rapp

In 1996 Rapp presents an interesting approach about automatic detection of tones. A F0 parameterization is conducted that uses optimization algorithms to fit a mathematical function (see F0 parameterization function in figure 3.11) to a given F0 contour. Before the F0 parameterization can be applied the waveform has to be segmented (as in Wightman & Ostendorf's approach presented above), which is done automatically by forced alignment of the known text. Since automatic segmentation is not flawless it remains unclear how the algorithm reacts upon faulty segment boundaries. After the automatic segmentation a two-syllable window is used as parameterization domain which is normalized in duration by linear scaling using -1 to 0 for the first, and 0 to 1 for the following syllable. Rapp's approach is inspired by Discourse Representation Theory (DRT) (Kamp & Reyle, 1993), a model about discourse semantics that "describes the interpretation of discourses as a dynamic two-level process." (Rapp, 1996, p. 2). Rapp intends to label only those intonational events "which are distinctive in the sense that one can assign them a function in the domain of discourse interpretation." (ibid. p. 2). The author uses the Stuttgart intonation model (Mayer, 1995) in his approach. This system has five standard pitch accents: a L*H (rise), a H*L (fall), a HH*L ("early peak"), a L*HL (rise-fall/"late peak"), and a H*M (stylized contour), and tries to integrate the phonological analysis of German intonation provided by Féry (1993) (see also p. 60) and the ToBI labeling conventions.

Rapp discusses some aspects regarding alignment between syllables and individual pitch accents and mentions situations where the two-syllable window might be insufficient for complete modeling. He also states the following: "As the algorithm optimizes locally, it is not guaranteed that it finds the global optimum, that is, the best phonetic parameterization. To avoid misparameterizations, an initial guess is computed by simple heuristics that should hopefully lie in the vicinity of the global optimum" (Rapp, 1996, p. 3).

Since there is no comparison of the algorithms performance with, for instance man-



$$f(t) = \alpha tanh(\beta(t-\gamma)) + \delta e^{-(\varepsilon(t-\zeta))^2} + \eta \qquad (function\,1)$$

Figure 3.11: Illustration of F0 parameterization function (cf. function 1) used in Rapp's approach. Original F0 values are depicted as circles. The parameterization function (solid line) is a result of: "tanh (dotted line), e^{-x^2}(dashdot line), constant (dashed line)" (Rapp, 1996, p. 3-4).

ually produced labels, it remains unclear how the performance of the algorithm should be estimated. Furthermore, since the approach needs a segmentation of the input speech the same critique as expressed for the Wightman & Ostendorf approach holds for Rapp's approach. The treatment of boundary tones is also not dealt with explicitly and it remains an open question how these are handled.

3.3.5 Ostendorf & Ross

Ostendorf & Ross (1997) present a "multi-level model for recognition of intonation labels". Before presenting their own model, they review a number of earlier approaches in this domain and group them in two classes: "those that model complete F0 contours and those that are a transformation of local F0 patterns and other cues given an utterance segmentation" (ibid., p.292). Under the first class they subsume the approaches of Ljolje & Fallside (1987); Butzberger et al. (1990); Chen & Withgott (1992); Jensen et al. (1993); Nakai et al. (1995); Geoffrois (1993); Hirai et al. (1995) and under the second class the models of ten Bosch (1993); Kompe et al. (1994) see section 3.3.7; Wightman & Ostendorf (1994) see section 3.3.2; and Campbell (1994) (cf. Ostendorf & Ross 1997, p. 292). The models belonging to the first class are said to work well for representing F0 contours but do not use durational cues which in turn are effectively used in the second class of models.

Ostendorf & Ross' own model is based on the assumption that "an utterance is a sequence of phrases, each of which are in turn realized as a sequence of syllable-level tone labels, which are in turn realized as a sequence of acoustic feature vectors (fundamental frequency and energy) depending in part on the segmental composition of the syllable" Ostendorf & Ross (1997, p. 291). They use a statistical approach to represent acoustic observations in the F0 and energy domain. Additionally they incorporate a stochastic segment model to account for observation sequences in syllables. To represent dependencies between acoustic features a hierarchy of levels (segment, syllable, phrase) is used. The model builds partly on parts of the Fujisaki model (accent filtering and superposition) but "includes additive Gaussian terms so that it is a probabilistic model" Ostendorf & Ross (1997, p. 292). The ToBI (see section 3.2.1) model was used as back end though, as it appears quite often in automatic approaches, the full label inventory was reduced into four types of accent labels (unaccented, high, downstepped high, low) and two intonation phrase boundary tones (L-L% and H-L% grouped as falling, and L-H% and H-H% as rising) and the three intermediate phrase accents (L-, H-, and !H-).

The authors emphasize that the inclusion of information at different time scales is an important advantage. Though this is certainly an interesting approach to look at, it has to be clarified what are the limits and the potential benefit of it. To account for segmental effects upon the F0 contour they refer to the recognition hypothesis (about the segmental content) which may then be incorporated into the detection process. However, it is not entirely clear how competing hypotheses about the segmental content would be accounted for.

To evaluate their system they compare the automatically annotated labels with manually established ones on basis of data from one speaker from the Boston University radio news corpus. The phone boundaries were known by forced alignment of the given word sequence and additionally manually labeled intermediate phrase boundaries were also known and simplified the recognition task. Therefore the results are not really representative. The labeling accuracy is estimated per syllable. 91% of the 3366 syllables manually labeled as unaccented were also recognized as unaccented; 7% of them as high and 2% as downstepped. 89% of the manually as high labeled syllables were labeled also high by the recognizer, 7% as unaccented, 3% as downstepped and 1% as low. 63% of the manually as low labled syllables were recognized as unaccented, 17% as high, 15% as downstepped and 5% as low. From the manually as downstepped labeled syllables were 25% recognized as unaccented, 39% as high, 35% as downstepped and 1% as low. These results show that the high accents are recognized fairly well but the downstepped and low ones are poorly. Regarding the intonational phrase boundary tones the recognition results are as follows: From the manually as falling labeled syllables 88% were recognized as falling 7% as rising. Manually as rising labeled syllables were recognized in 24% of the cases as falling and in 62% of the cases as rising. The latter result shows that there is still some amount of mismatches regarding this opposition of boundary tones.

Ostendorf and Ross' approach has an appealing idea, namely that it includes means to represent acoustic feature dependencies on several levels (segment, syllable, phrase). However, the complexity of the stochastic model is fairly large and nevertheless there remain quite a number of mismatches between the manual and automatically labeled test set especially regarding the low accents and the rising boundary tones as mentioned above.

3.3.6 MOMEL

MOMEL (MOdélisation de MELodie) was first proposed by Hirst (1980, 1983) and later implemented into a computer program by Hirst & Espesser (1993). MOMEL is an algorithm for the automatic modeling of F0 curves. The authors assume that the phonetic representation of a F0 curve is provided by a sequence of target points. MOMEL is intended to calculate these target points in a four stage process which includes: (1) preprocessing of F0, (2) estimation of target-candidates, (3) partition of candidates, (4) reduction of candidates. The first step includes the elimination of F0 values that are "more than a given ratio (typically 5%) higher than both their immediate neighbors [...]" (Hirst et al., 2000, p. 8). This step eliminates often one or two values at the beginning of voicing. The second step estimates target candidates by iteratively applying specific analysis methods within an analysis window of about 300 ms length. The result is one target value or a missing value for each original F0 value. The third step includes the partitioning of target points by the usage of another moving analysis window of about 200 ms length. In this window the average value of the targets in the first halve of it is compared to the average value in the second halve. The resulting "boundaries of the partition are then taken as those values which correspond to a local maximum for this distance and which is greater than the overall average value of the distances." (ibid., p. 8). The fourth and final step deletes further outlying values within each segment of the partition and calculates the mean value of the remaining values in each segment which is then the final estimate. MOMEL analyzes F0 contours, applies means to reduce microprosodic effects and generates target points in the F0 contour. These target points are then transformed to the INTSINT transcription system (see section 3.2.2). MOMEL includes some interesting approaches especially with regard to the elimination of microprosodic effects on the F0 contour.

Another approach entitled "Semi-automatic tagging of intonation in French spoken corpora" (Campione, 2001), also uses the MOMEL algorithm. The authors aim at a broad prosodic transcription for syntactic and pragmatic studies. Though it is not a fully automatic procedure, they argue that it is an approach "that automats critical steps in the labeling process and therefore reduces annotation time and improves the objectivity and coherence of the labels" (Campione, 2001, p. 90). The authors state that there is a lack of prosodically annotated corpora as a result of difficulties of prosodic transcriptions. The latter are time consuming and require phonetic competence of the annotator which is not common among syntax scholars. Addi-

tionally the subjective nature of prosodic labeling reduces the trustworthiness and needs even more time consuming control by other experts.

Campione and Véronis dispute the usefulness of the ToBI system because it is too detailed for syntactic studies. Their approach consists of 5 steps: (1) automatic detection of pauses in the speech signal, (2) stylisation of the F0 curve in order to eliminate microprosodic effects, (3) reduction of the stylized curve to a sequence of discrete symbols encoding the pitch movements, (4) orthographic transcription and synchronization to pauses and major pitch movements, and (5) filtering of the sequence of melodic elements and translation to the final prosodic labels (cf. Campione 2001, p. 92). Step 1 and 2 include automatic processing and manual correction, step 4 is purely manually, the rest automatically done. They use a system of 7 symbols without phonological meaning to describe the intonation contour of a given utterance: L+ large falling movement, L medium falling movement, L- small falling movement, S very small or null movement, H- small rising movement, H medium rising movement and H+ large rising movement. The stylisation of the F0 curve is provided by MOMEL. The output of the processing is a prosodically annotated text that is separated by paragraph marks for each prosodic unit. Prosodic events are marked at the end of each segment by one of three symbols: Start time of the respective unit is given at the beginning of the paragraph.

3.3.7 Verbmobil

In a series of papers[12] connected with the German Verbmobil[13] project (cf. Wahlster et al. 1997; Wahlster 2000) a prosody module was developed that was designed to improve the output of this automatic speech translation system. The authors claim that the system is "the world wide first and so far only complete speech understanding system, where prosody is really used [...]" (Batliner et al., 2000, p. 108).[14] The prosody module interacts with several other modules within the whole system, that is syntactic analysis, dialog processing, semantic construction, translation, speech synthesis, and provides a number of improvements in these modules. One of the main improvements is achieved in the classification of word boundaries, that is the decision whether a "full prosodic boundary" or one of the other three boundary classes (intermediate phrase boundary, normal word boundary, agrammatical boundary, e.g. hesitation) used in the system appeared after a word (cf. Nöth et al. 2000, p. 523).

The concept of the Verbmobil prosody module is guided by the domain it is used within, that is the speech translation of appointment scheduling dialogs. Therefore

[12] Cf., e.g. Kompe et al. 1995; Hess et al. 1997; Niemann et al. 1997; Batliner et al. 1999; Batliner et al. 2000; Buckow et al. 2000; Nöth et al. 2000; Batliner et al. 2001a.

[13] Verbmobil was intended to provide a speech-to-speech (e.g. German-Japanese and vice versa) translation for appointment scheduling dialogs.

[14] The use of prosodic information in automatic speech recognition (ASR) systems has already been proposed in 1980 by Lea; cf. also Vaissière 1988; Waibel 1988; Nöth 1991; Kompe 1997.

the prosody module takes both the output of the automatic word recognition module (the so called "word hypotheses graph (WHG)", ibid., p. 520) and the speech signal as input. This is motivated by the availability of the phoneme classes and the time-alignment. The output of the prosody module is a WHG with annotated probabilities for accent, clause boundary, and "sentence mood". Two basic classes of prosodic features are extracted: (a) acoustic features from the speech signal like F0, energy, and durational features which are provided by the output of the word recognizer, and (b) linguistic features provided by lexicon lookup, for instance syllable boundaries or position of lexical stress. The authors mention that it is still an open issue, especially for spontaneous speech, which prosodic features are necessary for the classification, and how they are connected with each other (cf. Nöth et al. 2000, p. 523). No attempt is undertaken to solve this question by phonetic or linguistic analysis but instead it is handed over to a statistical classifier. Therefore as "many relevant prosodic features as possible are extracted from different overlapping windows around the final syllable of a word or a word hypothesis" (Nöth et al., 2000, p. 522).[15] By doing so they end up with 276 features which consider a context of ± 2 words. Though the authors also present a study (Batliner et al., 1999) where they reduced this enormous feature set to 11 for boundaries and 6 for accents while simultaneously keeping the recognition rate in reasonable areas, it is questionable whether this large feature set can be motivated on linguistic reasons.

The reference point for the computation of the prosodic features is the end of a word. Among the features used are: duration for each syllable nucleus, syllable, and word; for each syllable and word the normalized (to mean F0) minimum and maximum of F0 and their position relative to the reference point, absolute and normalized maximum energy and their position relative to the reference point (for more detailed information about the set of features used see Nöth et al. 2000, p. 522-523). Though Kompe (1997, p. 191-193) lists the error rates of the F0 tracker used in the Verbmobil approach and also mentions that the "fine shape of the contour is not very informative with respect to accentuation and boundaries" (ibid., p. 193) there is no further notion of how the approach deals with faulty or microprosodically affected F0 values.

To train their statistic classifiers reference labels are needed which are provided by perceptually labeled boundary and accent classes (cf. Reyelt & Batliner, 1994). These classes include four different types of word-based boundary labels: B0: normal word boundary, B2: intermediate phrase boundary with weak intonational marking, B3: full boundary with strong intonational marking, and B9 "agrammatical" boundary, for instance hesitation or repair. The four labels for syllable based accents are: PA: primary accent; SA: secondary accent; EC: emphatic or contrastive accent; A0: any other syllable, not labeled explicitly (cf. Batliner et al., 1999, p. 2315). However, this set of prosodic labels is reduced to a two way distinction in both cases, that is whether there is or is not a boundary after a word

[15] The feature selection procedure is described in detail in Kießling, 1997.

or whether or not the word is accented. Three classes of "sentence mood" are distinguished: statement, question, and continuation rise.

The prosodic boundary labels seemed to be not sufficient which encouraged the authors to develop a new labeling scheme which they call "The Syntactic-Prosodic M-Labels" (cf. Nöth et al. 2000, p. 523 and Batliner et al. 1996, 1998). These labels (placed at the end of the relevant domain) mark different prosodic domains like "main/subordinate clause", "embedded sentence/phrase", "constituent, marked prosodically" (cf. Nöth et al. 2000, p. 523 ff) and are, among other things, motivated by their inclusion of labels specific for spontaneous speech. A total of nine classes is differentiated but reduced to the main three categories for the use in Verbmobil. Pattern recognition means are used to train models on basis of the manually labeled data. The average recognition rate for the classification of boundary vs. no boundary is 88.3% and the one for accented vs. unaccented word is 82.6%. These recognition rates could only be achieved when using the whole feature set. Though, when using only F0 features the recognition rate for accents dropped only slightly to 79.4%. In (Nöth et al., 2000, p. 525) it is stated that for boundary classification F0 and energy are the most important features and for accent classification F0 is most important and in contrast to the boundary classification more relevant than the energy features. Interestingly in a paper presented later the authors changed the order of relevance of these features towards the following hierarchy: duration, energy, pauses, F0 (cf. Batliner et al. 2001a).

The WHGs are subsequentely annotated with the probabilities for each prosodic class. These probabilities are then used by other modules in the Verbmobil system, that are: **Syntax**, **semantic construction**, **dialog processing**, **transfer**, and **speech synthesis**. In order to make the prosodic information annotated available to the **syntax module** a symbol for a clause boundary is introduced "at positions where either a M3 or a B3 boundary is expected" (Nöth et al., 2000, p. 527). The authors show that the number of different readings as well as the parse time is significantly reduced by the usage of these prosodic information. The **semantic construction** module uses information about accents for the interpretation of particles like "noch" *still* to disambiguate competing discourse representation structures. The module responsible for **dialog processing** has to identify dialog acts like greeting, confirmation of a place, etc. Astonishingly the dialog act recognition drops significantly (cf. Kompe 1997, p. 283) when using the automatically recognized segment boundaries. The latter lacks a detailed explanation except that there has been not enough training data. The **transfer module** involves the translation from German to English. Here accent information is used to disambiguate the interpretation of particles and sentence mood information is used for the distinction of questions and non-questions. Finally it was intended to adapt the **synthesized** output of the system to the voice characteristics of the original speaker. Though, the synthesis is only switched to a male or a female voice with respect to the input F0 contour. No details are stated about the reliability or benefit of this.

The prosody module in Verbmobil has shown that prosodic annotations can be of

great benefit for individual other modules of such a system, like syntax parsing or dialog processing, but there are still a number of open issues regarding the question how to integrate prosodic information successfully in existing or future ASR systems. The way prosodic information is acquired in the Verbmobil approach is by using information extracted from the raw speech signal like F0 and energy and simultaneously using the information provided by the automatic speech recognition module represented by word-hypothesis graphs. Enormous statistic processing is used to handle the input parameters as is expressed inter alia by the fairly large feature set (276) used. Even in some cases the intended statistical processing of the input data could not be conducted since the amount of time it needed prohibited their usage (cf. Nöth et al. 2000, p. 526 and p. 528). Despite the fact that Verbmobil showed a number of pros and cons in the usage of prosodic information in an automatic speech translation system, some of its conceptual aspects especially regarding the enormous statistical effort, are questionable.

3.3.8 ToBI Lite

Wightman et al. (2000) present an automatic prosody labeling approach which they call "ToBI Lite". It is intended to improve *unit selection*[16] speech synthesis quality. The authors claim that their system is more robust and simpler than standard EToBI (English ToBI). Interestingly they state also that the "reliability among labelers for some EToBI categories [Syrdal & McGory (2000), NB] was too low for successful training of an automatic prosody recognizer using the full EToBI system" Wightman et al. (2000, p. 71). They map bitonal pitch accents to "**" and other pitch accents to "*", and only edge tones marking major phrases are mapped to "%", unaccented syllables are marked with "0" (Wightman et al., 2000, p. 72).

The system needs an (automatic) segmentation of the input speech, that is phoneme, syllable and word boundaries are known. Beside that 24 "linguistically motivated acoustic features" (Wightman et al., 2000, p. 72) are used in their system, of which some are binary (e.g. stress, word-final, word-initial, schwa) and others continuous (e.g. normalized duration, maximum/average pitch ratio). To train the system a database of 860 utterances read by one female professional speaker was used that had been prosodically labeled by expert ToBI labelers and which included the collapsed categories mentioned above (**, *, %, 0). It remains unclear whether each individual labelers ratings were used or whether the individual ratings were collapsed into a generally agreed mean description.

[16] Unit selection synthesis is an approach started in the late 90ies (e.g. Hunt & Black, 1996) with the goal to overcome the limitations of concatenative speech synthesis. The latter needs concatenation points at every phoneme and therefore introduces segmental disfluencies which reduce the speech synthesis quality and are resulting in the lack of naturalness. Unit selection synthesis needs a large corpus of recorded speech and the goal is to minimise concatenation points by selecting the longest available unit from the corpus.

The systems performance was tested on a set of 42 utterances from the same corpus used for training but which were held out from the training set. Comparisons of the automatically set labels with both, collapsed manual EToBI labels, and to perceptually established ToBI Lite labels were conducted. Compared to collapsed EToBI labels accuracy was 83,5% of non-accented syllables and 84,9% of "**" syllables were correctly recognized. When collapsing both accent categories (** and *) recognition accuracy for prominent versus unaccented was 69,3%, with 16,5% incorrect labels. Phrase boundaries are correctly recognized 93,4% of the time, with 2% false labels. Comparison of automatically set labels with perceptually established ToBI Lite labels showed similar results except that recognition accuracy for prominent vs. unaccented syllables was higher (76,2% correct) but also with increased number of incorrect labels (19,1%). Additionally they show the usefulness of automatically established ToBI Lite labels for improving the quality of a unit selection (see footnote 16 on page 72) speech synthesis system and explain it partly by the greater consistency of automatic labeling.

Wightman *et al.'s* approach seems to be very promising for speeding up the annotation time of speech synthesis corpora and subsequently increasing perceived speech synthesis quality. However, they have not shown whether the same program may be used for different speakers (e.g. males) nor how the algorithm reacts to poor signal quality or different speaking styles (non-professional, conversational, etc.).

3.3.9 Other Approaches

Noguchi et al. (1999) present an approach about automatic labeling of prosody in Japanese. They state that the original J-ToBI system is insufficient for automatic labeling (cf. Campbell 1996) and propose therefore another inventory which consists of three tone sequences that form (including only one pitch accent which is H*+L), according to the authors, the legal set of Tokyo Japanese. Input to this approach are the F0 values and the segmentation into words including part of speech tags. Decision tree learning methods are used to implement the method which uses prosodic, syntactic, segmental and contextual features as input. A precision of 53% for the H*+L accents are given, having 34% deletions, 9.8% insertions and 10.9% substitutions. The sources for the errors seem to come from two sources: "(i) gentle slope or short duration in a rise/fall region and (ii) an unvoiced region which obscures potential rise/fall" (Noguchi et al., 1999, p. 4). The authors do not discuss problems of faulty or microprosodically affected F0 values but mention that they apply a 5 point median smoothing of the original F0 values.

Vereecken et al. (1998) present an approach for the automatic prosodic labeling of six languages. Input to the system is both the speech signal and its orthographic representation. As in a number of other approaches the segment boundaries of the input speech signal are known by the automatic labeler. The classification system includes four levels of prosodic boundary strength (PBS): "0 refers to ordinary word boundary, and values 1, 2 and 3 refer to weak, intermediate and strong

boundaries respectively. Phrasal prominence is labeled by assigning to each word a prominence (PROM) value between 0 and 9, with 0 being used for words which are not at all prominent and 9 being used for most prominent words" Vereecken et al. (1998). Similar to Batliner *et al.* (cf. section 3.3.7) they leave the decision about the most appropriate feature set to a statistical classifier. An analysis window of three words preceding and two words following the boundary to classify is used. For prominence detection only the word to be classified is analyzed since the authors state that it did not help to include acoustic features describing the words surrounding the one to be scored. Altogether about 200 features (both acoustic and linguistic) are used. The system is evaluated on six corpora consisting of about 1450 isolated sentences. The corpora represent the six languages Dutch, American English, French, German, Italian and Spanish. The corpora were manually labeled by one native speaker or near-native labeler. Results show that the inclusion of linguistic features improved the prosodic labeling performance significantly.

The system presented by Vereecken *et al.* labels prosodic boundary strength and word prominence on a gradual scale. Input to the system is the speech signal, a phonetic segmentation and linguistic annotations (like part-of-speech labels, word frequency, punctuation, stress, ...). An acoustic feature extractor plus a cascade of multi-layer perceptrons are used for the automatic classification.

3.3.10 Summary

In conclusion, existing approaches about automatic detection of prosodic cues may be broadly separated into two classes: the first class are represented by those approaches, which segment the incoming speech into phonemes, syllables or words before the intonational analysis starts (Wightman and Ostendorf; Rapp; Ostendorf & Ross; Verbmobil; ToBI Lite), one could call them *pre-segmentation* approaches. These approaches have the advantage to use information provided by the segmental content like stress positions or acoustic information like pre-final lengthening; the second class are the approaches without pre-segmentation of the incoming speech (Pierrehumbert; Taylor; MOMEL) and solely analysing the course of F0. ToBI Lite was used for the prosodic annotation of a unit selection corpus and the authors showed that the automatic annotations improved the perceived quality of the synthesis voice (Wightman et al., 2000).

One issue runs like a thread through the presented approaches, the separation of segmental and F0 tracking specific influences on the F0 contour from the potentially meaningful parts. This seems to be a fundamental problem in the automatic analysis of F0 contours since F0 movements resulting from such influences may easily be interchanged with meaningful units and subsequentely disturb the reliability of the automatic transcription. This aspect will be explicitly dealt with in section 5.1.

Another aspect involves the procedure of how to map from the (quasi-)continuous

F0 values to the discrete phonological elements in order to cover the full range of possible contours in the speakers set. Individual solutions have been proposed often by reducing the full set of labels towards a two-way distinction of high vs. low or rising vs. falling. For instance, the ToBI Lite approach revealed that the full (E)ToBI label inventory was used too inconsistently by human labelers in order to use it for successful training of an automatic procedure. Pierrehumbert uses her phonological model and the associated finite state grammar (cf. section 3.1.5) to generate the set of legal contours. Taylor develops his own system by mapping the three basic elements of rise, fall, and connection to a set of tone labels (though it is not fully specified). ToBI Lite reduces the full ToBI inventory to two pitch accents and one boundary tone. Verbmobil also uses its own set of prosody labels but practically uses only the two-way distinction accented vs. unaccented and boundary yes or no. Therefore, the reduction of a full set of prosodic labels towards a small but more reliably identified subset seems to be more promising for automatic detection purposes.

After the review of the most influential intonation models and the review of some approaches of automatic prosody recognition the next chapter will present a new approach of automatic prosody recognition - the ProsAlign (automatic prosodic aligner) model.

Chapter 4

ProsAlign - the Automatic Prosodic Aligner

The next three chapters will introduce the concept (Ch. 4) of ProsAlign – the automatic prosodic aligner, its implementation in a computer program (Ch. 5) and the evaluation of the program (Ch. 6). This chapter outlines the basic architecture of ProsAlign. First the question is addressed: what are the relevant acoustic features for pitch accents and boundary tones? Here qualitative definitions of the tones are presented together with the problems of quantifying them. Second the method of acoustic parameter assessment is explained. Both visual inspection and quantitative analysis of F0 contours were used to acquire concrete base values of acoustic features from a manually labeled corpus. Additionally in this part the importance and limits of manually labeled corpora are discussed. Third the results of the parameter assessment program are explained. And finally, in section four, the method to map phonological entities to acoustic features is described.

4.1 What are the Relevant Acoustic Features?

In phonological descriptions pitch accents and boundary tones are usually qualitatively described with acoustic features like maxima in the course of F0, often occurring, for instance along a H* pitch accent. However, the crux of such explanations appears when trying to formalize and to quantify descriptions like:

> "A pitch accent may be defined as a local feature of a pitch contour - usually but not invariably a *pitch change*, and often involving a local maximum or minimum - which signals that the syllable with which it is associated is prominent in the utterance." (Ladd, 1996, p. 45-46).

This definition states that it seems to be helpful to detect pitch changes and also maxima or minima but the restrictions "usually" and "not invariably" indicate that those features are not 100% reliable and that also other features might be important. Note that there is no hint about the exact amount of the pitch changes or the relative height of a maximum. Furthermore, there is nothing said about other acoustic features of a pitch accent except that they signal "that the syllable with which it is associated is prominent in the utterance". Since prominence is also signalled by relative energy changes across the spectrum it seems to be important to take those into account as well (cf Sluijter, 1995; Sluijter & Van Heuven, 1996).

The ToBI label guide (Beckman & Ayers, 1997, beginning of section 4.2) states that pitch accent tones have to be marked at every accented syllable and the "lack of pitch accent assignment for a syllable will be interpreted as meaning that the syllable is NOT accented".

The following definition of a H* pitch accent from the ToBI label guide (Beckman & Ayers, 1997, in section 4.2) provides further aspects:

> " 'peak accent' – an apparent tone target on the accented syllable which is in the upper part of the speaker's pitch range for the phrase. This includes tones in the middle of the pitch range, but precludes very low F0 targets."

Here the speakers pitch range is also an important criteria for the detection of pitch accents.[1] However, this would mean that there is an estimation of an individual speakers pitch range before the actual pitch accent assignment. This is difficult to achieve in an automatic procedure because it usually includes the inspection of several or longer examples from an individual speaker and the calculation of the upper and lower limits. Since this would impose severe restrictions on the application of the algorithm no attempt was undertaken in the present approach to account for that. Though, whenever there is a need to integrate the pitch range of a speaker, for instance, for improved recognition results, there could be a mechanism that adapts acoustic feature values in the F0 domain to the individual range of the speaker.

A more detailed description of a H* pitch accent is given in the GToBI label guide (Grice & Benzmüller 1997, see section 3.2.3) where the so called 'peak accent' is characterized by an upward movement of the pitch and forming the peak of a more flat increase (see two examples in figure 4.1). There is no low or high target before the accented syllable. This is the default accent that is used when there are no clear indications for other accents. The accented syllable sounds high. The upward movement is not as high as in the L+H* case (see an example in figure 4.2). This definition includes more detailed aspects regarding the amount of pitch

[1] See also the discussion of this topic on page 49 and in Ladd (1996, Ch. 7).

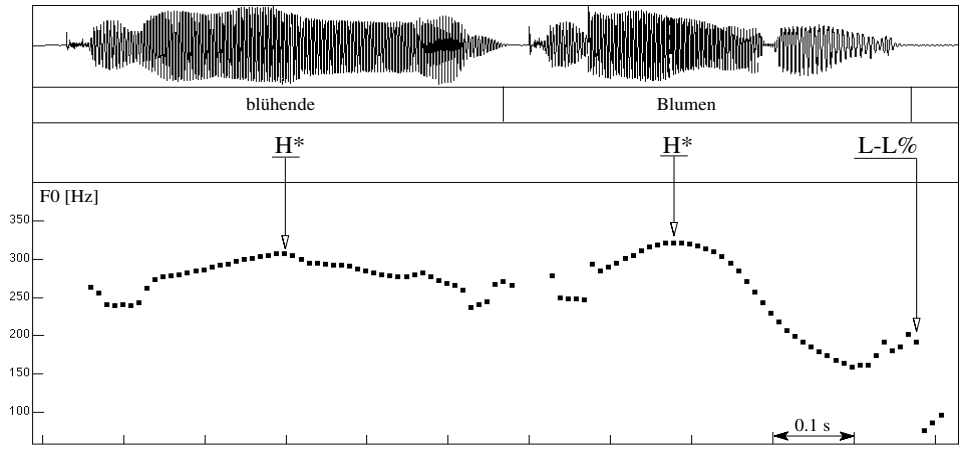


Figure 4.1: Example for two H* pitch peaks as labeled in the GToBI example "blumen2" (Grice & Benzmüller, 1997, taken from the training material). The text is "blühende Blumen" *blooming flowers*.

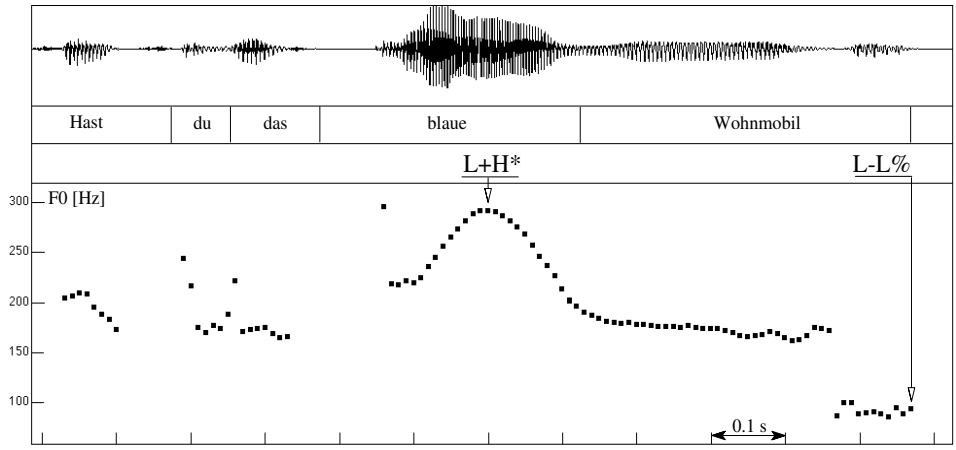


Figure 4.2: Example for a L+H* accent from the GToBI example "blaue" (Grice & Benzmüller, 1997, taken from the training material). The text is: "Hast du das blaue Wohnmobil" *Do you have the blue camper*.

increase (that is said to be not as high as in the L+H* case) and also about the preceding features where no low or high targets should be present. This indicates that the estimation of the amount of increases or decreases is significant. Moreover the 'history' of the F0 contour before the pitch accent is also important and might serve crucial decision criteria. Furthermore another crucial aspect is stated here namely that the "accented syllable sounds high". This subjective impression might be a serious criterium for human labelers but the concrete objective criteria remain nebulous.

A look at the qualitative definitions of pitch accents unveils problematic aspects when trying to integrate these into concrete numerical values. Therefore, it is essential first to get an overview about the type as well as the concrete numerical ranges of possible acoustic features. This is the subject of the following section.

4.2 Method of Parameter Assessment

In order to get concrete quantitative values for the underlying acoustic parameters a manually labeled corpus was taken as basis for an investigation. Though it is important to know that "Manually labeled speech corpora may not be sufficiently consistent for successful training or modeling for recognition or TTS systems" (Syrdal & McGory, 2000, p. 238) it provides a starting point for the estimation of acoustic features of the individual tones as well as their individual selectivity. Furthermore, to reduce inconsistencies between labelers, the GToBI training corpus was chosen because it includes prototypical examples of the individual tones and probably examples that represent a generally agreed mean description within the given framework. The GToBI corpus has the advantages of providing a reasonable number of examples for each individual pitch accent and boundary tone postulated in the underlying phonological model and also provides the acoustic material along with the prosodic label files (Grice & Benzmüller, 1997).[2] The inventory of pitch accents and boundary tones in the GToBI model is listed in table 4.1 and was discussed in section 3.2.3. For each of these pitch accents or boundary tones there are examples in the accompanying acoustic material of the GToBI training corpus. The quantitative analysis of these tones is described in the next section. Here the goal is to get an overview of the possible acoustic features and to assess reliable quantitative criteria for the individual tones that might later serve as selection parameters during the detection process.

Although the GToBI corpus is solely based on German speech material the basic method of the ProsAlign algorithm should be usable for any language. Possible adaptations in the acoustic features and necessary adaptations in the underlying

[2]The GToBI training corpus is available under http://www.coli.uni-sb.de/phonetik/projects/Tobi-/index_training.html

Pitch accents	*Boundary tones*
H*	L-
L+H*	L-L%
L*+H	L-H%
L*	H-
H+L*	H-L%
H+!H*	H-H%

Table 4.1: Inventory of pitch accents and boundary tones in the GToBI model (Grice & Benzmüller, 1997).

phonological inventory have to be made in order to adapt the algorithm to another language.

When analyzing acoustic features underlying manually labeled pitch accents and boundary tones, it becomes obvious on an early stage that the exclusive inspection of the course of F0 can only result in limited success with regard to automatic detection purposes. Because the F0 movements that are the indicators of individual pitch accents vary drastically and do often not provide sufficient selectivity. Moreover, those pitch movements can often not be separated from those that are not associated with pitch accents. Categorically different pitch accents like H* and L* can not simply be detected by searching for maxima (in the case of H*) or minima (in the case of L*) in the F0 contour. First of all these pitch accents are not always associated with maxima or minima. Second, the course of F0 does often not clearly distinguish them, and third there are different sized maxima and minima, that is having steeper or flatter increases or decreases before or after or a larger or smaller number of voiced neighbors that are not always associated with pitch accents. Without a clear definition of these parameters (and probably others as well) one can expect only limited detection success.

However, when one starts to define threshold values, for instance regarding the amount of increase in F0 for individual pitch accents, it soon will appear that there is on the one hand often a considerable amount of overlap in the search criteria to get a reasonable coverage and on the other hand a very poor recognition rate when the criteria is defined too restrictive. In addition, F0 movements consisting of faulty F0 values or microprosodically affected ones (see footnote on page 36 and section 5.1) might erroneously be taken as pitch accent indicating F0 movements. Therefore, the separation of those effects from linguistically meaningful parts within a F0 contour is an important aspect for an automatic detection procedure. With respect to this problem, the voicing parameter plays an important role, since it allows the determination of the location of an individual F0 value within a voiced part. In addition, knowing that F0 values are most often erroneous or microprosodically affected up to 5 periods from the beginning or end of voicing, the voicing parameter

could help to decide (together with other parameters) whether a F0 value is likely to be faulty or not.[3]

Based on the observations from the manually labeled pitch accents and boundary tones it became clear that the synchronous course of amplitude and its representation in the so called RMS amplitude is an important additional criteria to the course of F0 and may provide sufficient selectivity for an automatic detection purpose.[4] The course of RMS gives information about increasing and decreasing amplitude values as well as the relative height of maxima and minima in it and although it does not provide absolutely reliable information about it (since there is no segmental analysis), it provides important features about onsets and offsets and centres of vowels, syllables or words and subsequently of intonation phrases.[5]

The RMS amplitude (**R**oot **M**ean **S**quare amplitude) of a stretch of n samples that is said to be a rough estimate of perceived loudness (Reetz 1999, p. 19) is calculated as follows:

$$RMS\ amplitude = \sqrt{\frac{Sum\ of\ all\ squared\ elongations}{Number\ of\ elongations}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

In a concrete implementation this formula is applied within a predefined analysis frame of the speech signal. For example the *get_f0* program from the ESPS/waves-tools calculates RMS values "based on a 30 ms hanning window [...]" (Talkin & Lin, 1997, p. 1).[6]

As a consequence of the visual inspection of the possible acoustic features of pitch accents and boundary tones the following three parameters were taken as baseline in the present approach:[7]

[3] "The probability that an individual period will be markedly erratic from the trend is highest at a point up to five cycles after the onset or before the offset of voicing." (Laver, 1994, p. 453). See also Viswanathan & Russel, 1984.

[4] Interestingly Batliner et al., 1999 note that F0 features are not more important than energy or duration features in their evaluation of prosodic features for pitch accent and boundary classification.

[5] Relying solely on RMS features is of course not fully sufficient for those purposes but provides nevertheless important clues for it. The importance of the RMS feature is also reduced when there are strong background noises or other disturbing influences on the RMS contour.

[6] See also the description of *get_f0* in Talkin (1995). A 'hanning-window' is a specific type of analysis window also called 'cosine window' since it uses a cosine function ($w(i) = 0.5 + 0.5 * cos(\frac{2\pi i}{N})$) and is used to focus the calculations made in the central part of the analysis window and simultaneously putting less weight on the edges of it to reduce the influences of sudden jumps at these edges (see reference before and Reetz 1999, p. 72).

[7] Actually a fourth parameter is not listed explicitly here because it is a inherent feature of the mentioned parameters, that are durational parameters like the duration of increases or decreases in F0 or RMS; or the duration of voicing before pitch accent location, etc. "Duration" here does not include phoneme, syllable, or word duration since these units are not recognized in the presented approach.

1. F0,
2. voicing,
3. RMS amplitude.

Since voicing is a necessary requirement for the extraction of F0, it is possible to subsume it under the F0 parameter. However, since separate listing makes the criteria more transparent, it was kept an individual class. Other parameters like formants or phoneme durations would introduce another source of possible errors in the recognition step and since phoneme identity is not yet easily and reliably recognizable automatically, no attempt was undertaken to detect segment identity or exact segment boundaries.

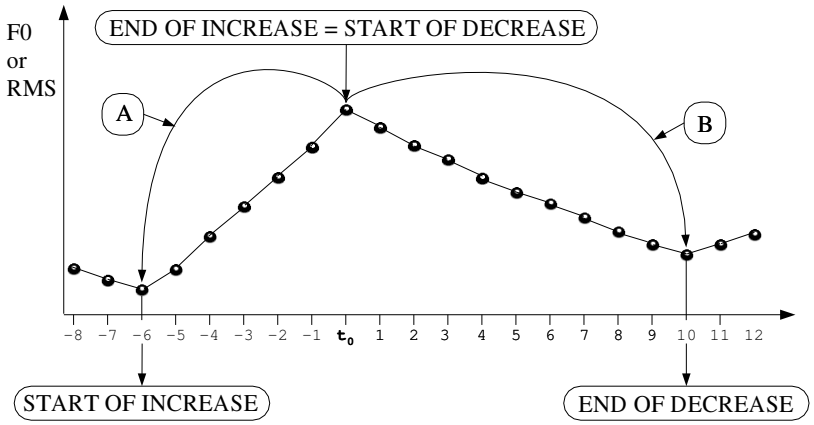
A parameter acquisition program was designed which aimed to acquire quantitative criteria for the subsequent implementation of the automatic detection program. But what exactly should be covered within the acoustic parameters? A first approximation towards an answer to this question was provided by the visual inspection of the acoustic features F0, RMS, voicing and duration, combined with the auditory control for several instances of each individual tone. The visual inspection of the course of F0 some distance around the tone location allowed us to think about possible strategies for capturing the F0 movements. The simultaneous auditory control served a further clue and though it remains an open question how to integrate the latter, especially without knowing the segmental content, it was basically used for the estimation of the relative importance of possible features.

Several possibilities are conceiveable to capture the acoustic features of tones, for instance, duration of a voiced stretch, duration of F0 increase, amount of F0 increase, etc. Associated questions are: Where to start or stop the duration measurement of a F0 increase? How to calculate the amount of F0 increase (relative or absolute)? Are there correlations between the three parameters and how to account for them? What temporal domain should be covered? As a starting point and as a result of the visual and auditory inspection of a number of tones[8] it was decided to analyze the manually labeled pitch accents and boundary tones in the GToBI corpus with respect to the following criteria:

- duration of increasing and decreasing parts of F0 and RMS before and after,
- amount of increase since the start of increase and amount of decrease before the end of decrease (see figure 4.3),

[8]Batliner et al. (1999) present another method of finding the most efficient parameter set for automatic classification of prosodic events by using linear discriminant analysis (LDA) to minimize the number of features while simultaneously preventing too much loss in classification performance. They start with 276 features and reduce this set to 11 for boundaries and 6 for accents. Among the features for accent classification are: lower energy after accent location; more energy variation at accent location, F0 is falling before and rising at accent location.

- duration of voiced or voiceless parts before or after point t_0 (see explanation below).



$$\text{(A)} \quad \text{AMOUNT of INCREASE} = \frac{\text{Value at the start of increase}}{\text{Value at the end of increase}}$$

$$\text{(B)} \quad \text{AMOUNT of DECREASE} = \frac{\text{Value at the start of decrease}}{\text{Value at the end of decrease}}$$

Figure 4.3: Idealized illustration of F0 or RMS track for showing the method of amount estimation.

Each manually labeled tone was analyzed with respect to the above-mentioned parameters within an interval of $\pm$400 ms around its labeled position. The decision for this time window was based on a first inspection of pitch accents and boundary tones and seemed to be a reasonable analysis frame as to cover enough contextual material for the selection of acoustic features.

The decision to use the linear Hertz frequency scale was based on the knowledge that the transformations of the F0 values into a logarithmic scale did not seem to be of significant influence for the approach chosen here. Although some researchers have explicitly chosen the logarithmic scale (semitones) because it represents human perception of frequency (cf. the IPO model in section 3.1.1 and 't Hart & Cohen 1973; Silverman 1987), the transformation of frequency values from a linear scale to a logarithmic one does not solve the problems faced with when modelling the F0–phonology interface, and it was decided to stay with the linear Hertz frequency values (cf. Taylor 1994, p. 85-86).

The parameter analysis should result in the identification of perceptually important F0 movements as well as in the differentiation of these movements from perceptu-

ally unimportant F0 movements, and the contribution of the other parameters for this percept. At this point it should also become clearer whether there are specific combinations of the three parameters under investigation that are able to characterize pitch accents and boundary tones at the acoustic level.

The individual parameters are summarized in table 4.2 and illustrated in figure 4.4. Since values for each parameter are given every 10 ms, these are always calculated starting from a virtual point t_0. Therefore, the range of increase or decrease estimations is calculated in frames from starting point t_0. Increase or decrease estimation was calculated by comparing neighboring values and adding up the number of fulfilled cases, in the case of an increase before the number of items where $F0_{t_x} > F0_{t_{x-1}}$ $(x = 0-40)$ was and in the case of a decrease after the number of items where $F0_{t_x} > F0_{t_{x+1}} (x = 0-40)$ was. The possible range of values that an individual parameter may take on is also given. Since the number of increasing or decreasing values in the parameters F0 and RMS very rarely exceeds 20 values (which means 200 ms) it is restricted to this limit. The values for the amount estimations are either between 0 and 1 for cases where the ratio of a smaller value to a larger one is made (e.g. AF0inA = minimum F0 value compared to F0 value at the end of an increase afterwards) or >1 whenever the numerator is larger than the denominator (e.g. AF0inB = maximum F0 value compared to F0 values at the begin of an increase before, see figure 5.5 on page 113 in the following chapter). For more detailed aspects regarding the parameter acquisition program see section 5.3. In the following section the results of the parameter assessment program are presented and discussed.

4.3 Results

The results of the parameter assessment program were used for further statistical processing in order to formulate adequate detection criteria. An illustration of an excerpt from the output is given in table 4.3). From each of the acoustic features the mean, median, standard deviation, as well as the minimal and maximal value were calculated with a standard statistic program (StarOffice5.2 Calc). The tables will present the median because it is usually a more representative measurement for highly variable data, meaning that it is less affected by single extreme values. Although the standard deviation is connected with the calculation of the mean, it is listed in order to show the variability existing in the individual parameters. A summary of the results for all the pitch accents in the GToBI training material is presented in table 4.4 and for all the boundary tones in table 4.5. The results will now be discussed in more detail, first for the pitch accents and second for the boundary tones.

Parameter	Range	Name
Voicing		
nr of voiced values before t_0	0-40 [frames]	*VoicB*
nr of voiceless values before t_0	0-40 [frames]	*VoilB*
nr of voiced values after t_0	0-40 [frames]	*VoicA*
nr of voiceless values after t_0	0-40 [frames]	*VoilA*
F0		
nr of increasing F0 before t_0	0-20 [frames]	*F0inB*
nr of decreasing F0 before t_0	0-20 [frames]	*F0deB*
nr of decreasing F0 after t_0	0-20 [frames]	*F0inA*
nr of increasing F0 after t_0	0-20 [frames]	*F0deA*
amount of F0 increase before t_0	1-$\star$	*AF0inB*
amount of F0 decrease after t_0	1-$\star$	*AF0deA*
amount of F0 decrease before t_0	0-1	*AF0deB*
amount of F0 increase after t_0	0-1	*AF0inA*
RMS		
nr of increasing RMS values before t_0	0-20 [frames]	*RMinB*
nr of decreasing RMS values before t_0	0-20 [frames]	*RMdeB*
nr of increasing RMS values after t_0	0-20 [frames]	*RMinA*
nr of decreasing RMS values after t_0	0-20 [frames]	*RMdeA*
amount of increase in RMS before t_0	1-$\star$	*ARMinB*
amount of decrease in RMS after t_0	1-$\star$	*ARMdeA*
amount of decrease in RMS before t_0	0-1	*ARMdeB*
amount of increase in RMS after t_0	0-1	*ARMinA*

Table 4.2: Table of parameters in the parameter acquisition program.

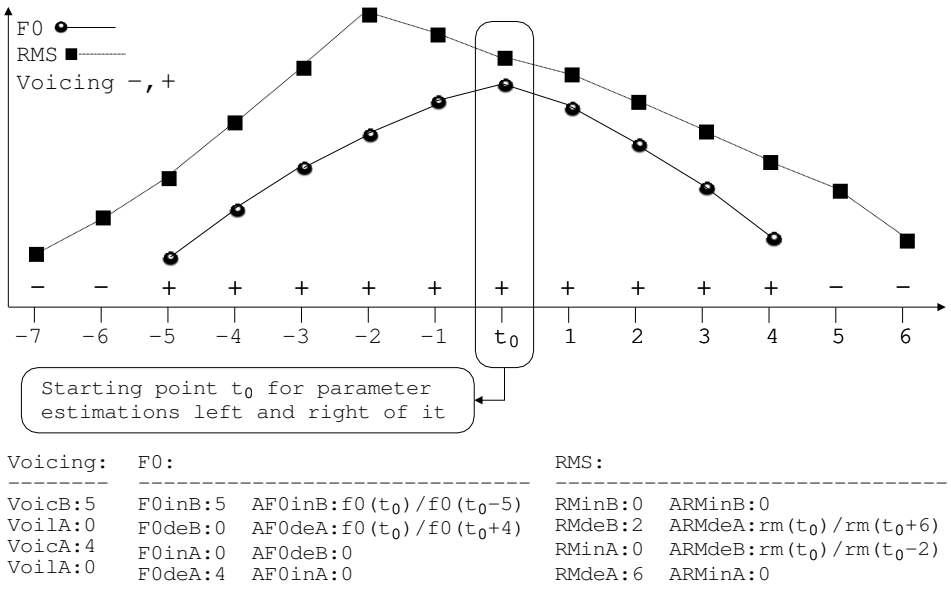


Figure 4.4: Idealized illustration of F0 and RMS tracks for showing the 20 acoustic parameters chosen for the classification of pitch accents and boundary tones. The value of each individual parameter is presented underneath the chart relative to starting point t_0 as depicted.

4.3.1 Results for Pitch Accents

The first impression when looking at the combined results is the large variability in most of the parameters that is expressed by a large standard deviation. The results in the voicing domain indicate that most of the pitch accents are not labeled at the beginning or end of voiced parts (indicated by VoilA/B = 0). Usually more than 7 values (= 70 ms) before and more than 10 values (= 100 ms) after a pitch accent are voiced.

The results in the F0 domain show for instance that the number of continuously increasing F0 before (F0inB) as well as the amount of increase before (AF0inB) is, as expected, larger for L+H*- than for H*-accents. It also becomes obvious that high pitch accents are not always marked at maxima (that should show up as F0inB > 0 and F0deA > 0) in the F0 track nor are low pitch accents always marked at minima (that should show up as F0deB > 0 and F0inA > 0). High pitch accents seem to be marked preferably slightly before the maximum as indicated by the high number of increasing values before and the small number of increasing values after. The downstepped accent H+!H* forms a separate case marked preferably in a decreasing F0 part (F0deB > 0 and F0deA > 0). Also the L*+H and the H+L* cases are preferably marked in decreasing F0 parts whereas the L* seems to be generally marked in increasing F0 phases (F0inB = 1 and F0inA = 9).

Name	Tone	Time [sec]	F0 [Hz]	RMS	VoicB	VoilB	VoicA	VoilA	…
august	H*	1.40	247	1415	5	0	30	0	...
blumen2	H*	0.31	306	4100	23	0	29	0	...
blumen2	H*	0.81	320	3468	16	0	33	0	...
dina4	H*	2.07	245	2749	9	0	40	0	...
g-essen	H*	0.68	270	1711	40	0	4	0	...
goldmine	H*	0.46	232	9336	5	0	5	0	...
...	...	...	...	...	...	...	...	...	...

Table 4.3: Segment of the results of the parameter analysis program for some of the H* pitch accents in the GToBI training corpus. The column entitled 'Time' provides the point in time when the individual H* occurs calculated in seconds from the beginning of the file. The following two columns show the individual values of F0 and RMS at this point, and the remaining columns show the numbers of continuously voiced or voiceless values before and after (VoicB/VoilB, VoicA/VoilA), relative to this point. Altogether there were 20 parameters extracted, 4 in the voicing domain, 8 in the F0 domain, and 8 in the RMS domain.

Not surprisingly the number of increasing pre-accent F0 values is highest for L+H* (F0inB: 8), followed by H* (3) and interestingly the median value for the L* cases is 1 which indicates that they are not always labeled at minima in the F0 track, although close to one. Similar observations can be made in case of the two high pitch accents H* and L+H* where the median number of increasing post-accent F0 values is in both cases 1 and therefore indicates that they are also not always labeled at maxima, however at least close to one. The L* cases seem to be characterized by a large number of increasing F0 values afterwards (F0inA: 9) which is significantly higher than for all the other accents including the L*+H cases (F0inA: 0) where one would expect it.

As control the number of increasing and decreasing F0 values before and after as calculated by the program were visually checked by inspecting the F0 track. Here it became clear that the simple criteria, for instance $F0_{t_x} > F0_{t_{x-1}}$ (for increase before) is insufficient, because there are often cases where F0 values do not fulfill the criteria. Nevertheless, the visual inspection clearly shows an increase or decrease. Therefore, the estimation of increases and decreases has to be improved in order to come closer to the ability of humans when visually inspecting F0 tracks. Humans are able to smooth small deviations from a general trendline and these deviations may vary themself from a single outlying value up to several ones also including interruptions by voiceless parts. The human ability of visual integration was already picked out as a central theme in Pierrehumbert (1983, see section 3.3.1). However, there was no obvious solution for the integration of this ability into a pitch accent detection algorithm.

The amount of increase before (AF0inB) is as expected highest for the L+H* cases (1.17) followed by the H* cases (1.04) and small for the L* cases (1.01). The

Tone	**H***		**L+H***		**H+!H***		**L***		**L*+H**		**H+L***	
nr items	51		25		7		11		7		6	
	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD
VOICING												
VoicB	7	(13)	15	(12)	11	(9)	17	(13)	9	(8)	23	(12)
VoilB	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
VoicA	13	(13)	19	(13)	40	(14)	24	(12)	10	(15)	10	(15)
VoilA	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
F0												
F0inB	3	(4)	8	(4)	0	(1)	1	(3)	0	(3)	0	(.5)
F0deB	0	(1)	0	(0)	1	(4)	0	(3)	1	(5)	2	(2)
F0deA	0	(4)	0	(5)	1	(3)	0	(1)	1	(1)	5	(2)
F0inA	1	(3)	1	(2)	0	(1)	9	(8)	0	(2)	0	(0)
AF0inB	1.04	(.5)	1.17	(.1)	0	(.5)	1.01	(.5)	0	(.6)	0	(.5)
AF0deA	0	(.6)	0	(.6)	1.01	(.7)	0	(.5)	1	(.5)	1.09	(0)
AF0deB	0	(.4)	0	(0)	0.75	(.4)	0	(.4)	0.79	(.5)	0.81	(.4)
AF0inA	0.93	(.5)	0.77	(.5)	0	(.4)	0.51	(.3)	0	(.5)	0	(0)
RMS												
RMinB	0	(4)	0	(3)	0	(3)	1	(7)	1	(5)	4	(1)
RMdeB	1	(2)	1	(2)	1	(3)	0	(2)	0	(1)	0	(0)
RMinA	0	(1)	0	(.3)	0	(.4)	1	(4)	0	(1)	1	(.5)
RMdeA	4	(4)	5	(4)	6	(4)	0	(2)	1	(3)	0	(3)
ARMinB	0	(66)	0	(6)	0	(8)	1.09	(8)	1.01	(8)	3.65	(4)
ARMdeA	3.04	(14)	3.37	(.13)	3.52	(58)	0	(5)	1.03	(13)	0	(2)
ARMdeB	0.52	(.4)	0.72	(.4)	0.73	(.5)	0	(.4)	0	(.4)	0	(9)
ARMinA	0	(.3)	0	(.3)	0	(.4)	0.4	(.4)	0	(.5)	0.92	(.4)

Table 4.4: Median values (Md) and standard deviations (SD, in brackets) for the mentioned acoustic features of pitch accents in the GToBI training material.

amount of increase afterwards is fairly high for the L* cases (AF0inA: 0.51) indicating steep increases afterwards. In the case of the other two low pitch accents the decrease before is around 0.8 and the decrease seems to continue in the case of the H+L* (AF0deA: 1.09) whereas it is only 1 for the L*+H cases where one would expect a steeper increase.

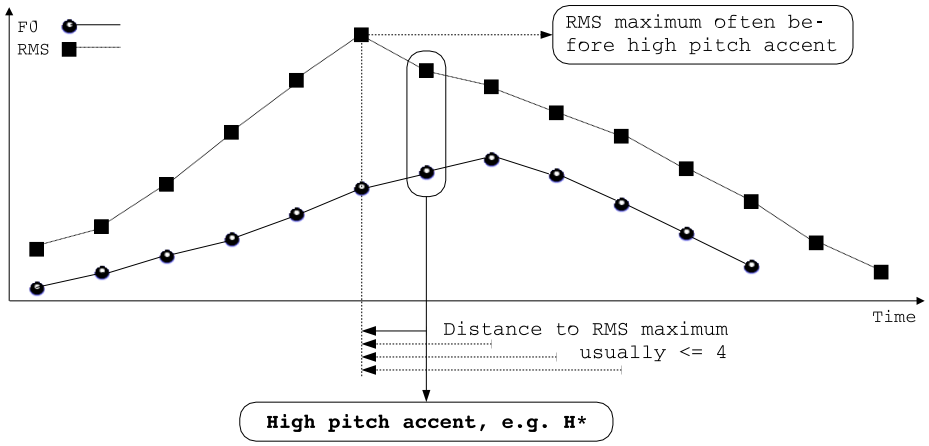


Figure 4.5: Illustration of the alignment of high pitch accent locations and RMS maxima. Often the distance of pitch accent label and preceding RMS maxima is <= 4 values.

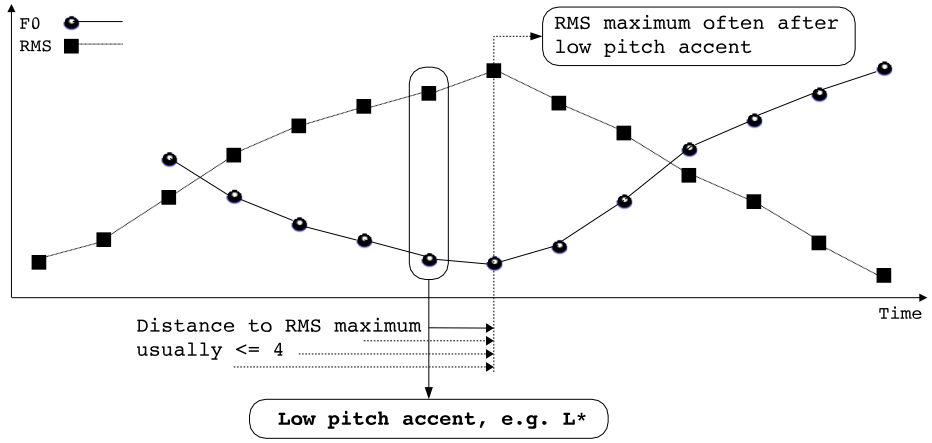


Figure 4.6: Illustration of the alignment of low pitch accent locations and RMS maxima. Often the distance of pitch accent label and following RMS maxima is <= 4 values.

The results in the RMS domain include some interesting findings: Maxima in the RMS track are usually only 1 value before the position where high pitch accents are marked (see figure 4.5) and low pitch accents are usually also marked fairly close to a maximum in the course of RMS, but slightly before (see figure 4.6). When

checking these relations for the individual pitch accents the following relations appear: in 61% of the high pitch accents (H*, L+H*, H+!H*) is RMdeB>0 and in 71% of the low pitch accents (L*, L*+H, H+L*) is RMinB>0. Of course these relations are tendencies, but no strict limits. These findings and the small standard deviations in these results indicate fairly reliable search criteria in this domain.

Since high pitch accents are usually marked slightly after the RMS maximum the amount of decrease before is large (ARMdeB: about 0.7) as well as the amount of decrease after (ARMdeA: about 3), whereas the low pitch accents are usually marked slightly before a RMS maximum and therefore the amount of increase before (ARMinB) is >1 and the amount of increase after is large (ARMinA: 0.4 for L*; 0.92 for H+L*). However, the L*+H cases deviate from this regularity, indicating that they are marked exactly at a RMS maximum (ARMinB > 1 and ARMdeA > 1).

The low pitch accents show a more diverse picture. Where one would expect fairly high values for the amount of decrease before and fairly low ones for the amount of increase after as can be observed in the L* cases (ARMinA: 0.4) and in the H+L* case (ARMinA: 0.92) this does not hold for the L*+H cases.

The analysis of the acoustic parameters of pitch accents revealed the following results: (i) high pitch accents are neither marked always at maxima in the F0 track, nor are low pitch accents marked always at minima. However, both cases are usually close to maxima (for high accents) or minima (for low accents) in F0; (ii) interestingly, the relative position of RMS maxima to pitch accent labels is usually remarkably close. That is, high accents occur slightly (<= 40 ms) after RMS maxima and low pitch accents occur slightly before RMS maxima (<= 50 ms); (iii) all pitch accents are usually labeled more than 60 ms after beginning of voicing or more than 90 ms before the end of voicing; (iv) estimations of the amount of F0 increase before revealed that H* and L+H* accents are usually labeled in increasing F0 parts and that the size of increase is higher for L+H* accents, as one would expect from their definition; L* accents show fairly steep F0 increases after label position; (v) the amount of RMS decrease after label position of high pitch accents is more than 3, that is the RMS value at label position is 3 times larger than at the end of the following fall; low pitch accents are marked by RMS increases before label position, though to a lesser extend than the high pitch accents.

The visual control of automatic F0 increase or decrease estimations revealed that the simple selection criterium is insufficient for reliable estimates. Only a single deviation from the estimation criterium (e.g. $F0_{t_x} > F0_{t_{x+1}}$) could result in a incorrect judgment, since often slight deviations from this criterium occur that are nonetheless perfect increases or decreases. Since this estimation is directly related to the fundamental question in intonation research, which F0 movements are perceptually important and which are not, it needs an adequate solution. However, it is by no means obvious how to set up more adequate detection criteria for these purposes. This difficulty represents a tightrope walk between smoothing some of the devia-

tions on the one side and leaving (perceptually) important movements unsmoothed on the other side. In order to strike a happy medium an approach is presented in section 5.3.2 to account for this aspect.

4.3.2 Results for Boundary Tones

The results of the parameter assessment program for the boundary tones in the GToBI corpus are summarized in table 4.5. First the four intonation phrase boundary tones will be discussed and afterwards the two intermediate phrase tones. In the case of H-L% boundary tones there are only two items which is not sufficient for representative statistics, however, they were listed for reasons of completeness.

Intonation phrase boundary tones

Since boundary tones are labeled at the end of an intonation- or intermediate-phrase, they are often at locations that are voiceless and are usually followed by pauses of differently sized durations. This explains why all the median values in the F0 estimation domain are equal to zero.

Besides, one expects to get a larger number of voiced values before the boundary since there is usually speech with its typical amplitude variations and its successive changes of voiced and voiceless parts. This is reflected in the results with a large number of voiceless values after the boundaries (median values for VoilA: L-L%: 40; L-H%: 29, H-H%: 40, and H-L%: 25). Here, it has to be mentioned that the algorithm counts values beyond the beginning and end of the file always as "0", that is whenever the analysis window overlaps the limits of the given input file, missing values are filled with zeros.

In all intonation phrase boundary tones the number of decreasing RMS values before (RMdeB) is greater than or equal to 2, as one expects, since usually the amplitude decreases in the final segment towards the end of the phrase. The estimation of the amount of decrease before label position reflects this as well (ARMdeB) which is usually < 0,25. It also seems that the boundary locations are often in falling parts just one value before a local minimum in the course of RMS amplitude indicated by the small number of decreasing values after (RMdeA: about 1). To summarize the results for the boundary tones: (i) Boundary tones are most often labeled in voiceless parts; (ii) pauses after intonation phrases are often marked by long (>250 ms) voiceless stretches; (iii) RMS usually decreases towards the boundary tone.

Intermediate Phrase Boundary Tones

The intermediate phrase boundary tones show a different picture since they are much more often labeled at points that are voiced and do not have such long pauses

Tone	**L-L%**		**L-H%**		**H-H%**		**H-L%**		**L-**		**H-**	
nr items	27		5		8		2		10		25	
	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD
VOICING												
VoicB	0	(16)	0	(0)	18	(19)	0	(0)	0	(13)	24	(17)
VoilB	2	(7)	13	(3)	4	(8)	17	(3)	2	(12)	0	(6)
VoicA	0	(3)	0	(0)	0	(13)	0	(0)	0	(15)	4	(15)
VoilA	40	(18)	29	(11)	40	(16)	25	(16)	3	(12)	0	(8)
F0												
F0inB	0	(.5)	0	(0)	0	(0)	0	(0)	0	(1)	0	(.8)
F0deB	0	(.8)	0	(0)	0	(2)	0	(0)	0	(3)	0	(2)
F0deA	0	(.8)	0	(0)	0	(0)	0	(0)	0	(2)	0	(1)
F0inA	0	(.3)	0	(0)	0	(.7)	0	(0)	0	(0)	0	(2)
A1F0inB	0	(.3)	0	(0)	0	(0)	0	(0)	0	(.8)	0	(.5)
A1F0deA	0	(.6)	0	(0)	0	(0)	0	(0)	0	(.6)	0	(.5)
A1F0deB	0	(.4)	0	(0)	0	(.4)	0	(0)	0	(.2)	0	(.3)
A1F0inA	0	(.3)	0	(0)	0	(.3)	0	(0)	0	(0)	0	(.4)
F0smB	0	(9)	0	(0)	0	(9)	0	(0)	0	(6)	1	(13)
F0smA	0	(2)	0	(0)	0	(12)	0	(0)	0	(9)	0	(13)
RMS												
RMinB	0	(.9)	.5	(.5)	0	(.7)	0	(0)	0	(.8)	1	(1)
RMdeB	3	(5)	2	(2)	4	(6)	4	(2)	1	(6)	0	(4)
RMinA	0	(.6)	0	(0)	0	(1)	.5	(.5)	.5	(2)	1	(3)
RMdeA	1	(1)	1	(0)	1	(1)	0	(0)	.5	(2)	0	(1)
ARMinB	0	(.6)	0.5	(.5)	0	(.6)	0	(0)	0	(3)	1	(1)
ARMdeA	1.09	(2)	1.17	(.1)	1.05	(.9)	0	(0)	0.57	(1)	0	(1)
ARMdeB	0.13	(.4)	0.06	(0)	0.05	(.2)	0.26	(.1)	0.1	(.3)	0	(.3)
ARMinA	0	(.5)	0	(0)	0	(.4)	0.49	(.4)	0.03	(.2)	0.09	(.4)

Table 4.5: Median values (Md) and standard deviations (SD, in brackets) for the acoustic features of the boundary tones in the GToBI training material.

afterwards as in the case of the intonation phrase boundary tones. Therefore the number of voiced items before, especially in the case of the H- tones, is fairly large (median value of VoicB: 24). Nearly all of the other parameters do not include sufficient selectivity in this case. However, with respect to the intermediate phrase boundary tones one cannot expect to get consistent acoustic feature values since these tones are often set according to labeling conventions that are not based on the existence of acoustic features at a particular point in time but on the sequence of tones and their wellformedness.

Despite this last aspect, the results for the boundary tones indicate that the chosen parameters are not sufficient to get enough criteria and subsequently satisfactory selectivity for their detection. Since intonation phrase boundary tones are often marked by relatively long pauses afterwards (often > 250 ms) criteria have to be integrated that are able to detect such cases.

4.3.3 Conclusion

In conclusion the results of this analysis show that the acoustic parameters analyzed provide a first approximation but are by no means sufficient for purposes of automatic detection of pitch accents and boundary tones. The large variability as well as the sometimes overlapping parameters indicate a small selectivity. Furthermore, the sometimes simple selection criteria have been shown to be deficient, as in the cases of F0 increase or decrease estimation before or after pitch accents. Here more appropriate criteria have to be developed in order to get better judgments in this respect and subsequently more reliable selection criteria. Beside the problems with the F0 track estimation the criteria for the RMS track indicated fairly reliable that pitch accents are usually marked close to a maximum in the synchronous RMS track. This finding promises to get more selectivity for the decision whether an individual F0 movement might be considered as a possible pitch accent candidate or not.

All these drawbacks in the detection of acoustic features reveal the structural limitation of a pure bottom-up procedure and leads us towards the motivation for the 2-step approach promoted in this thesis. Another structuring level is introduced following the detection of acoustic features, namely the phonological mapping procedure. In the latter, the final decision for a particluar tone at a specific position takes place. Therefore, before introducing the improved selection criteria, the phonological mapping process will be explained next.

4.4 Phonological Mapping

The association of pitch accents with acoustic features needs a mapping algorithm that connects both sides of these domains. Since phonological entities are discrete and symbolic on the one side, whereas acoustic features are continuous and

numeric on the other side, it is a one-to-many mapping. Several different constellations of numeric values in the acoustic features may be subsumed under one pitch accent. The phonological mapping procedure has to decide what are the allowable variations in the acoustic feature values in order to assign an individual pitch accent or boundary tone to a set of acoustic features. Furthermore, the algorithm has to check whether specific features are present and possibly apply weights to individual features. For instance, whether there are 10 or 15 voiced items before or after might be less important than the number of continuously increasing F0 values before a H* pitch accent.

One method to get pitch accents or boundary tones from acoustic features could be the following: for each single tone concrete numerical threshold values for each of the acoustic features are defined. That means, a tone can be selected when *n* criteria are met or either *m* criteria or *l* other criteria. Implicitly the threshold values and the combinatorics of the individual features has to provide sufficient selectivity in order to ensure that not one and the same acoustic feature set will be assigned to two or more tones. This method would directly choose or reject tones from the acoustic feature values of an individual point in time. However, whenever a single criterion is not fulfilled, the individual point in time will be ruled out, that is, if there were a criterion that stated that a H* pitch accent must have at least two increasing F0 values before, every F0 value that does not fall into this range would be ruled out. Such a selection procedure is too restrictive. In fact, it became clear during a first study with this method that too many tones are missed, since the large variability in the acoustic feature values is often not covered by the threshold values.

Therefore, another method was developed to select pitch accent and boundary tone candidates from the acoustic feature values. This method first defines a 'fingerprint' for each individual pitch accent and boundary tone in terms of acoustic feature values. Each tone is defined with concrete numerical range values for each single acoustic feature and those range values are not directly taken as decision criteria, but are integrated in a scoring system that takes the individual importance of the features into account. Positive points are given for feature values that support the existence of an individual tone, negative points are distributed when they do not. Finally, all points are added up and the resulting score will be used to select tone candidates (see task-flow diagram in figure 4.7). Threshold values can be defined that select candidates that have a high score and deselect candidates that have a low score.

This procedure postpones the decision for a specific tone and first gets an overview of the acoustic features spectrum. The overview is expressed in a score that is used to select tone candidates. Therefore, this method leaves more space for possible candidates and does not already rule them out at this preselection level. In a final decision step, only one tone will be selected from a possible list of candidates at a given point in time, but this choice is additionally based on sequence restrictions of the pitch accents and boundary tones. In this connection, the pitch accents im-

mediately before and after are checked both for their identity as well as their score. First of all, there could be several identical tones in a row. In this case, the algorithm selects the one with the largest score under the assumption that the higher this number, the greater will be the correspondence with the acoustic feature values of an individual tone.

Second, there could be different types of pitch accents immediately following each other. In this case the score is once again taken as decision criterion for the one or the other tone. However, since there may for instance be a high pitch accent marked just a few milliseconds before an intonation phrase boundary, it could also be deleted when the score for the boundary is high and the boundary tone could be transformed to one ending in a high tone when it was not already detected. Therefore, deletions or transformations of tones are possible in this algorithm based mainly on the score, but also on sequential aspects.

Another conceivable possibility in the phonological mapping process could be the application of wellformedness restrictions according to the postulations in the underlying phonological model. However, in order to test the validity of the acoustic feature set on basis of the existing procedure and simultaneously not introducing another source of possible errors, it was decided not to include such wellformedness conditions at those stage.

After the basic architecture of the ProsAlign program has been laid out the following chapter will now describe the implementation of the method in a computer program.

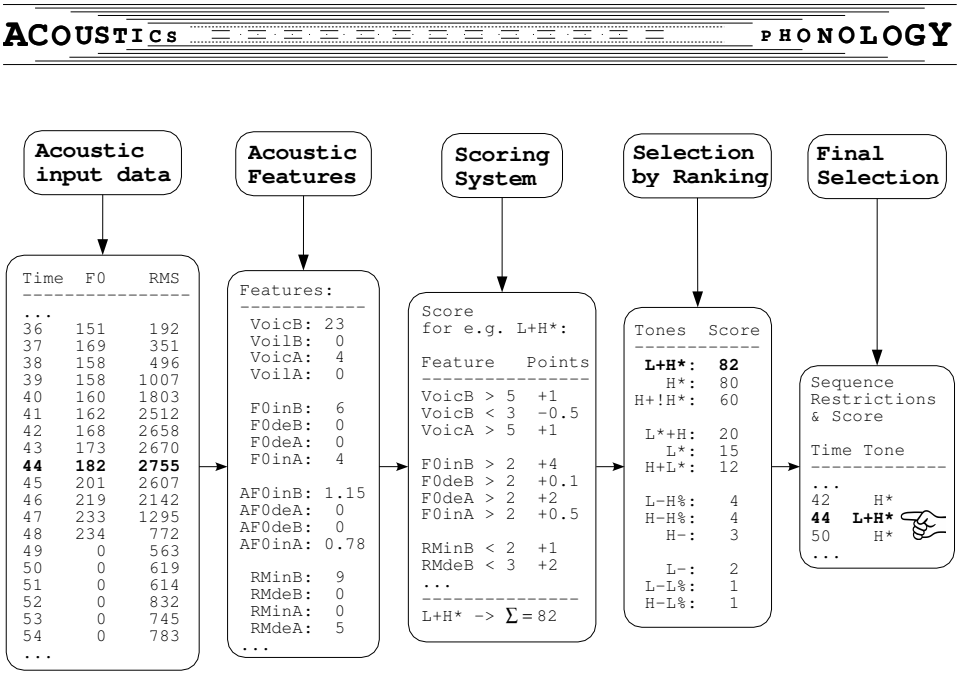


Figure 4.7: Task flow from the continuous parameters of F0, voicing and RMS over the estimation of acoustic features up to the phonological mapping procedure that evaluates the appropriateness of feature combinations for the individual pitch accents and assigns scores. The final selection includes the consideration of both sequence restrictions and score.

Chapter 5

Implementation of the Model

The goal of the work presented here is to set up an explicit model to describe the recognition of prosody. This means to model the acoustics – phonology interface. The basic architecture of this model is depicted in figure 5.1 and was laid out in chapter 4 before, regarding the phonological mapping process especially in section 4.4. This chapter describes the implementation of the model in a computer program. The handling of acoustic variability as well as the identification of potentially meaningful F0 movements are addressed.

The input to the program is the acoustic speech signal. It contains the speaker individual characteristics of sound pressure changes and is a continuous signal that is digitized with a certain sampling rate. Since the underlying hypothesis in the approach presented here is that not only the course of F0 is important for the automatic detection of prosodic events but also the synchronous course of energy (represented as RMS amplitude), the first processing step is the extraction of the three acoustic parameters fundamental frequency (F0), root-mean-square amplitude (RMS), and voicing. These parameters are calculated stepwise every 10 ms by the *get_f0* program (version 1.14) from the ESPS/waves+ tools.[1] The step size of 10 ms is justified by a good time resolution for catching enough detail in the changing parameters and a reasonable calculation time. The parameters and especially the F0 contour include segmental (microprosodic) effects as well as erroneously estimated values resulting from poor signal quality or problems specific to the method of F0 extraction.

The calculated acoustic parameters F0, RMS, and voicing are then analyzed by a feature extraction process. Here features like the duration and the amount of increases and decreases in the course of F0 and RMS are calculated. Other features are the position of local maxima and minima in the course of F0 and RMS. These parameters are calculated by framewise comparison, that is, by comparing the F0 value at a point t_0 with its immediate neighboring F0 value at a point $t_0 - 1$, al-

[1] The *get_f0* program is described in detail in Talkin (1995).

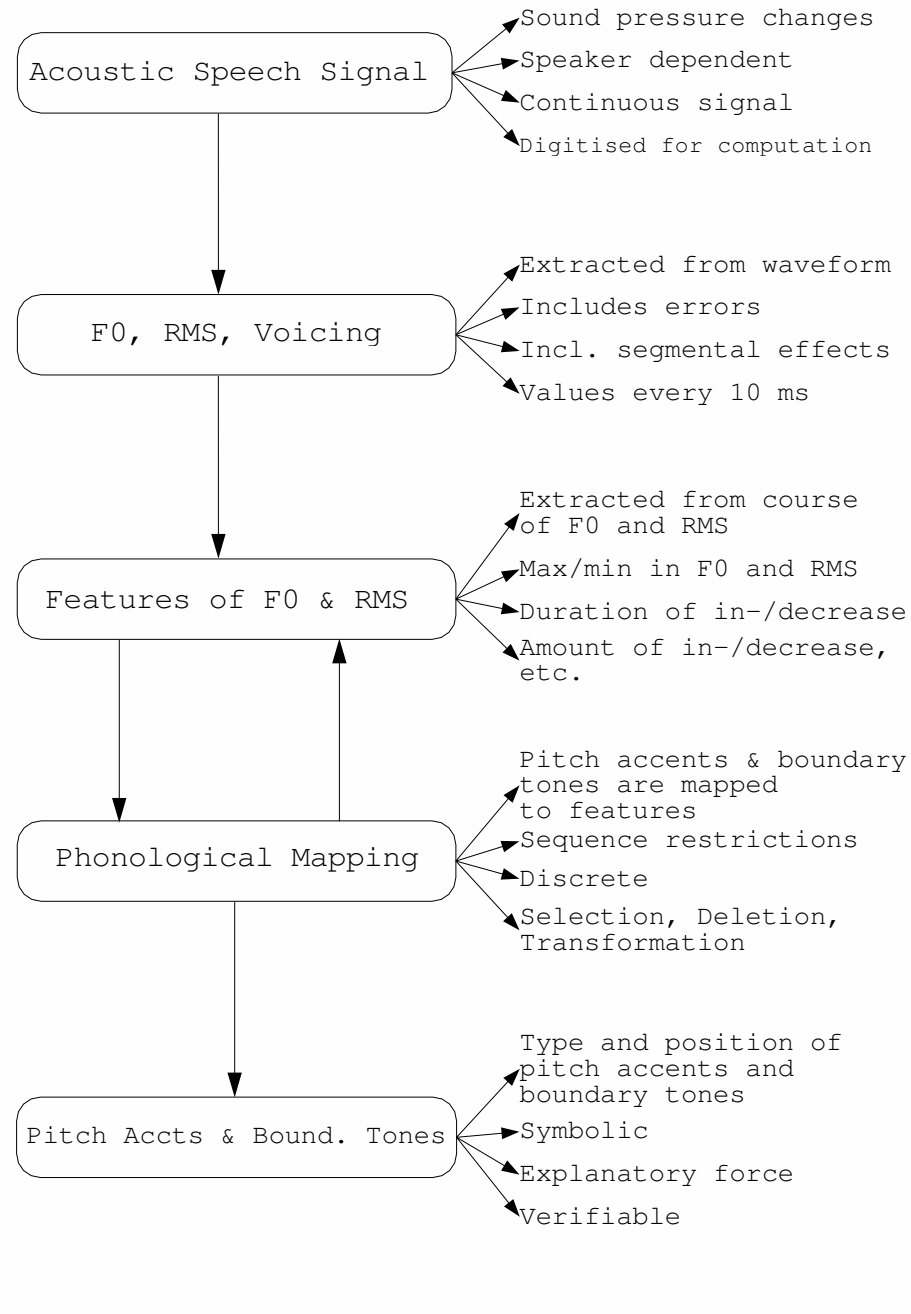


Figure 5.1: Outline of the model underlying the automatic prosodic aligner ProsAlign (see text for description).

though comparisons also include comparisons of more distant frames, for instance RMS value at point t_0 is compared to RMS value at point $t_0 - 10$. Since a number of effects in the course of F0 are not relevant for the preception of prosodic events, means are developed to separate those cases from the potentially meaningful movements. The latter forms an important part in the series of processing steps within the program.

The extracted acoustic features are subsequently fed into the phonological mapping module. The latter maps combinations of acoustic features to pitch accents and boundary tones. These prosodic events are defined by a phonological model about the underlying structure of intonation in a given language. The pitch accents and boundary tones are discrete and are defined in terms of the position they have to be assigned to and the type of F0 movement associated with them. Since the phonological model structures the incoming acoustic features by phonological rules, this is a clear top-down processing which is represented by the arrow pointing from the phonological mapping box to the "Features of F0 & RMS" box in figure 5.1. The phonological mapping process is implemented in a scoring system that assigns positive scores to feature constellations which are supporting the existence of an individual prosodic event and negative scores to constellations that do not. Finally, the score is used together with rules that may select, deselect or transform tones to produce the output, that is the series of pitch accents and boundary tones. Type and position in time of these symbolic categories is then available and describes the underlying prosodic structure in the given speech file.

The design of the program is driven by the underlying phonological model of intonation presented in section 3.2.3 (i.e. ToBI and its German implementation GToBI). The detection procedure therefore, explicitly incorporates a mapping algorithm of phonological elements to acoustic features representing the structuring influence of abstract entities in the classification of highly variable acoustic input data.

Moreover, the sequence of processing steps is intended to represent the difference between **recognition**, the ability to differentiate sound signals, and **perception**, the ability to associate sound signals with meaningful units. The recognition part is represented in the first two steps, whereas the third step represents the perception part. However, there is no clear separation of these two areas since the second step could be interpreted as already using pre-selection criteria extracted from the acoustic analysis of pitch accents.

It is known that the calculation of F0 by pitch tracking algorithms introduces some errors (see 5.1). Possible sources of errors are incorrect voicing detection, pitch halving and doubling errors, and incorrect F0 extraction due to poor signal characteristics (background noise, laryngealizations, creaky voice, breathy voice, etc.). Also segmental effects occur on the course of F0 that are introduced by the coarticulation of vowels and consonants. Sharp rises or falls after stops in the first periods of voicing are typical. Such influences may disrupt the smoothness of the course of F0 but have not been shown to be of importance for the perception of the more

general structure of intonation.[2] However, these artifacts have to be taken into account when the extracted F0 values are taken as the basis for selecting prosodic events. The differentiation of faulty or microprosodically affected F0 movements from potentially meaningful ones is an integrated part in the detection algorithm. One of the strategies chosen for the separation of those influences is provided by putting less weight on the beginnings and ends of voiced periods in the selection processes.

Since the separation of faulty or microprosodically affected F0 values from potentially meaningful ones is an important factor in this program it will be discussed in the following section.

5.1 Faulty or Microprosodically Affected F0 Values

Because the output of F0 trackers is known to be faulty (above and Reetz 1996 and Hess 1983) and includes a number of segmental effects known as microprosodic perturbations, these cases have to be separated from the other essential F0 values. ProsAlign performs this separation within a scoring system (cf. section 4.4) that is designed to estimate F0 values by giving negative points for potentially faulty or microprosodically affected values. However, the first step here is a detailed analysis of F0 tracks and the possibilities of identifying faulty or microprosodically affected F0 values. In a first approximation, this identification mechanism checks for absolute differences in adjacent F0 values, since faulty as well as microprosodically affected F0 values are often characterized by extreme jumps (>25 Hz) from one to the next F0 value. However, since this detection criteria are sometimes not reliable or sufficient, additionally the synchronous course of voicing as well as the course of RMS is taken into consideration.

What is a faulty F0 value?

A faulty F0 value could be defined as a F0 value given by the pitch tracker that has no obvious basis in the corresponding waveform, that is measuring the pitch manually indicates that actually a different F0 value is present. However, sometimes this definition does not cover cases where the manual measurement is unclear but the perceptual impression clearly indicates a different pitch than the pitch tracker calculated. This happens sometimes with laryngealizations at the end of phrases usually indicating a fall to the speakers bottom pitch range but is often calculated

[2] Sharply rising or falling F0 movements of limited duration (<50 ms) at the boundaries from or to voiced or unvoiced stops do not result in the perception of pitch changes, because their duration is usually below the critical duration (about 6 cycles) that is necessary to perceive a certain pitch height. However, that does not mean that they are not perceptible, they are used as acoustic cues for the identification of the accompanying consonant (Helfrich 1985, p. 89 and Haggard et al. 1981).

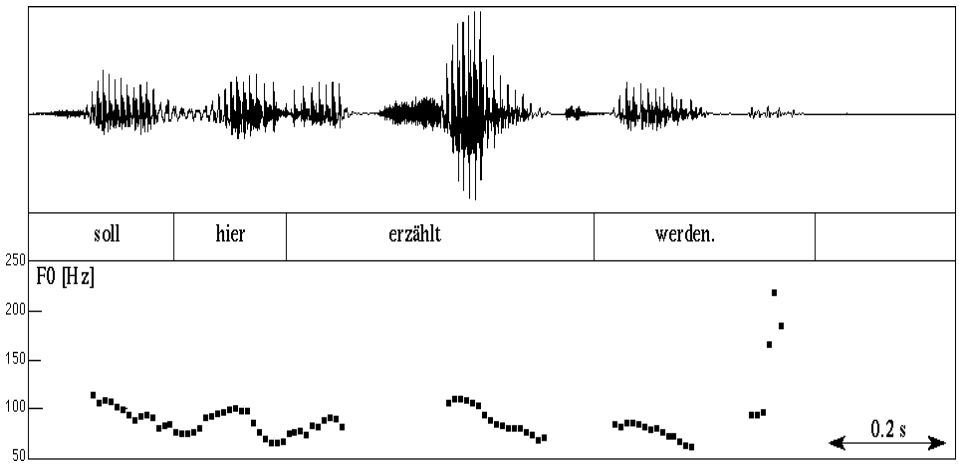


Figure 5.2: Picture of waveform and original F0 track of the phrase "[...] soll hier erzählt werden" *[...] should be told here* showing the effect of laryngealization on the F0 contour at the end of the phrase.

with very high F0 values by the pitch tracker. Because no direct detection of laryngealizations were possible, it was decided to do this indirectly by the time synchronous inspection of the RMS curve. Typical cases of faulty F0 values at the end of phrases were characterized by strong changes from one to the next pitch value (>25 Hz) accompanied by a small RMS amplitude as compared to the previous voiced part. A typical example is phrase final [ən] which is usually produced as a syllabic [n̩] in German. Such an example is illustrated in figure 5.2.

Another problem in pitch tracks are local (i.e. within 1-3 frames) outlying F0 values. Here the visual inspection instantaneously ignores such a jump but the maxima detector could be misled. Local outlying F0 values are usually characterized by sudden jumps in the frequency value from one frame to the next. A difference greater than 25 Hz along with a smooth course of F0 before and after (about 40 ms) is a fairly reliable indication that the value is erroneous. Moreover, the course of RMS amplitude is also an additional indicator of the reasonableness of a frequency value because local outliers appear most often in parts where the RMS amplitude is suddenly changing or is strongly reduced compared to parts before or after. Therefore the joined observation of F0 and RMS course is used as decision criterium for separating regular pitch movements from faulty ones. It has to be mentioned that this method works very well in many cases, but fails in some cases, for instance when speakers change to their falsetto voice and produce sharp final rises. The F0 fault detection may well recognize such cases as erroneous and could mistakenly assign negative scoring to it. However, since the score is a result of several features there could be still a chance for such a case to become a pitch accent candidate, which shows the advantage of the chosen method.

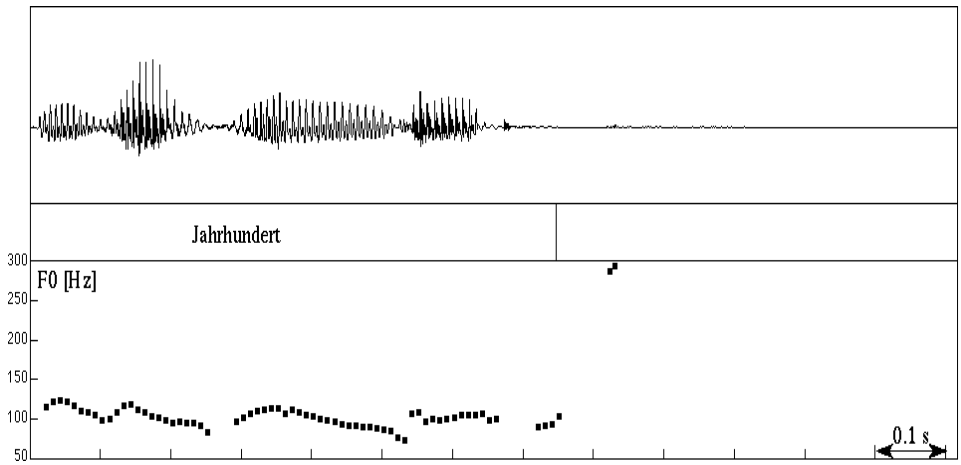


Figure 5.3: Waveform and F0 track of the phrase final word "Jahrhundert" *century* with two outlying F0 values afterwards.

Figure 5.3 shows another phenomenon in F0 tracks: outlying F0 values in parts that are most of the time voiceless and have nearly zero level amplitude. Here are two F0 values slightly below the 300 Hz limit line after the end of the word "Jahrhundert" *century* in the original F0 track. The amplitude of the waveform is extremely low. These two F0 values come actually right after an intonation phrase and are part of a breath pause. The perceptual impression is just a breath pause, not a short period of high pitch, and therefore these two F0 values can be neglected. Such cases receive negative scoring points from the scoring algorithm as a result of these contextual acoustic features. Of course, most often such cases can be identified when looking at their short duration (about 40 ms or 1-4 successive F0 values).

Frequency halving or doubling

Frequency halving or doubling is the effect that the pitch tracker shows erroneously the half or double of the actual F0 due to a incorrect estimation of the size of the glottal pulses. Measuring the glottal pulses by hand can confirm that. Although efforts were made to detect such cases, no satisfactory solution was found and therefore no correction for this effect was applied. However, trying to correct such errors is quite dangerous, because

> "[...] it is [...] not obvious whether such a jump is an error of the algorithm or whether it is a quality of the speech signal. Octave jumps are part of normal communication and can even be part of the linguistic inventory of a language (Huber, 1988). Smoothing pitch

> contours by hand or automatically in order to remove seemingly obvious octave jumps might remove a characteristic of the speech signal or the linguistic message and consequently is dangerous." (Reetz, 1996, p. 135-136).

Transitions from unvoiced to voiced parts and vice versa

A typical microprosodic effect at transitions from unvoiced to voiced parts (e.g. a sequence of unvoiced stop – vowel) are a sharply falling F0 track into the more stable part of the vowel (see e.g. Lehiste 1970, p. 68 ff). The algorithm therefore puts less weight on the beginnings and ends of voiced parts particularly when there are large differences between neighboring F0 values. A negative score might be given to separate these cases from potentially meaningful high–low F0 movements. The actual integration of these effects into the algorithm is provided in the scoring system. Negative scores are used to reduce the total score of a potentially microprosodically affected or a faulty F0 value. The individual weights of these scores were designed with respect to the positive scores in order to allow the reduction of the total score below the selection threshold.

5.2 Detection of Acoustic Features

After the possible error sources in the F0 tracks have been addressed, the next step is the reliable detection of acoustic features. This step comes right after the extraction of F0, RMS and voicing from the input speech signal and is an important process, since the underlying hypothesis of the approach presented here assumes that the acoustic features are the mediators between the continuous stream of data in the speech signal and the discrete phonological entities. As already mentioned in section 4.1 and 4.2, the following set of acoustic features was taken as a baseline in the present approach:

- duration of increasing or decreasing parts of F0 and RMS,
- amount of increase since the start of increase and amount of decrease before the end of decrease,
- duration of voiced and voiceless part before and after point t_0.

Since this set (as shown in the last chapter) does not provide a good coverage of the acoustic phenomena additional parameters were introduced:

- number of voiced or voiceless values before and after point t_0 without continuation restriction, that is the counting does not require uninterrupted voiced or voiceless parts,

- better estimations of increases or decreases in the course of F0, possibly coming close to the human ability of visually integrating small outlying values in a general increase or decrease,
- acoustic features that allow the detection of faulty or microprosodically affected F0 values,
- features that allow the detection of boundary tones and accompanying pauses.

In the following section all the parameters in the acoustic feature domain are listed and explained in detail. Also the method of parameter assessment for the acoustic features is presented followed by a discussion of its results.

5.3 Acquisition of Quantitative Criteria

Since the analysis of acoustic features of pitch accents and boundary tones in section 4.1 showed that the chosen parameters were not sufficient for a good coverage of the phenomena in the F0, RMS, and voicing domains, a second analysis was conducted. Further parameters were included, for instance a better estimation of F0 increases or decreases that allows a limited number of outlying values of a general increase or decrease.

The following three sections present the acoustic features first in the voicing domain (see section 5.3.1), then in the F0 domain (see section 5.3.2) and finally in the RMS domain (5.3.3). A program was developed that extracted numerical values of these parameters for all the pitch accents and boundary tones in a manually labeled corpus. The results of the program will be discussed later along the consequences for the overall design of the automatic detection algorithm.

5.3.1 Parameters in the Voicing Domain

Voicing is a fundamental feature used in phonetic classification and refers to the vibration of the vocal folds during articulation. A sound produced with vibrating vocal cords is called voiced, for instance [a, e, z] while those produced without such vibrations are called voiceless or unvoiced, for instance [t, h, f]. Here the voicing information is extracted from the speech signal by, roughly speaking, searching for the quasi-periodically repeating glottal pulses as opposed to the non-periodical signals in voiceless stops or fricatives (cf. Hess 1983). When signal characteristics are good, there are usually no problems with this measurement. However, one has to mention that poor signal quality, background noise or other influences (creaky voice, laryngealizations, etc.) may affect the voicing detection. For instance, part of the speech signal might not be marked as voiced although there were actually

vibrating vocal folds during its production, or a stretch of speech is marked as voiced although the original signal was not produced with vibrating vocal folds.

For the approach presented here voicing estimation is very important since it is bound to the estimation of F0, *that is per definitionem* F0 is only represented in voiced parts of speech signals. Since it is well known (cf. Laver 1994, p. 453) that F0 values are most often erroneous up to 5 cycles from the beginning or before the end of a voiced stretch this serves important information about potentially faulty or microprosodically affected F0 values. However, this does not imply that there cannot be a pitch accent at the beginning or at the end of a voiced part. When other acoustic features are present that support the presence of a pitch accent then it is likely for example that a F0 value at the end of a voiced phase becomes a L+H* pitch accent candidate.

Moreover, it is important to know how long a voiced part in a speech signal is, or expressed in another way: it is important to look at the distance of a potential pitch accent candidate from the beginning and end of the voiced part it is located in. Very short stretches of voicing (<5 cycles) are less reliable locations of pitch accents than longer stretches as a result of the above mentioned errors at beginnings and ends of voiced parts. However, these short stretches are still possible locations of pitch accents since many short vowels fall into this time domain. Additionally, it is helpful to know the duration of the unvoiced stretches before and after a voiced phase. Since these are important features for the estimation of short term interruptions of voicing or longer possibly pause-like breaks. Therefore, the following 16 parameters in the voicing domain are introduced (see table 5.1 on page 121). All parameters are computed for every frame, which is virtually represented by the term t_0. The decision for a $\pm$400 ms analysis window was based on the first feature analysis (cf. section 4.1) and proved to be a reasonable analysis frame in order to cover enough contextual material for the selection of acoustic features and the subsequent mapping of tones.

Number of continuously voiced/voiceless values before or after point t_0:

Parameter	Range	Name
nr of continuously voiced values before t_0	0-40	*VoicB*
nr of continuously voicel. values before t_0	0-40	*VoilB*
nr of continuously voiced values after t_0	0-40	*VoicA*
nr of continuously voicel. values after t_0	0-40	*VoilA*

These parameters calculate the number of uninterrupted voiced or voiceless values in time frames before and after point t_0. Since voiced values are represented with a F0 value larger than 0 all values are counted that are larger than 0 before or after

the value under inspection without a single interruption by an unvoiced value that is represented with a 0. The other way around the number of uninterrupted voiceless values before and after is calculated. Measurement values may range from 0-40 frames. The algorithm counts values beyond beginning and end of the file always as "0" (= voiceless). Often speech files are cut off close to intonation phrase boundaries, which therefore increases the number of voiceless values towards the end of the file. However, this does not necessarily result in a boundary tone label at the end of each file, since also other parameters have to be present as well.

Number of voiced values before or after point t_0 without continuation control:

Parameter	Range	Name
nr of voiced values 50 ms before t_0	0-5	*Voic5B*
nr of voiced values 100 ms before t_0	0-10	*Voic10B*
nr of voiced values 160 ms before t_0	0-16	*Voic16B*
nr of voiced values 230 ms before t_0	0-23	*Voic23B*
nr of voiced values 310 ms before t_0	0-31	*Voic31B*
nr of voiced values 400 ms before t_0	0-40	*Voic40B*
nr of voiced values 50 ms after t_0	0-5	*Voil5A*
nr of voiced values 100 ms after t_0	0-10	*Voil10A*
nr of voiced values 160 ms after t_0	0-16	*Voil16A*
nr of voiced values 230 ms after t_0	0-23	*Voil23A*
nr of voiced values 310 ms after t_0	0-31	*Voil31A*
nr of voiced values 400 ms after t_0	0-40	*Voil40A*

These parameters calculate the number of voiced items in the interval represented by the number in it, that is *Voic5B* calculates the number of F0 values that are larger than 0 in an interval ranging from item t_{0-1} up to item t_{0-5}, that is 5 values before the value under inspection. In this connection it is important that there is no continuation control, so it is possible that voiced values are separated by short or long unvoiced stretches, but nevertheless they are counted when they are in the given interval. This is different to the measurement above, where only uninterrupted stretches of either voiced or voiceless parts were counted. The sizes of the intervals were chosen because they showed a good coverage of the 400 ms intervals before and after t_0. These parameters give estimations for the change of voicing before and after the value under inspection.

The number of voiceless values within the intervals is derived by subtracting the number of voiced items from the length of the interval, for instance number of voiceless values 50 ms before $= 5 - Voic5B$.

5.3.2 Parameters in the F0 Domain

In an algorithm that detects pitch accents automatically, the course of F0 is naturally very important. Although this may at first sight seem easy, one of the difficult aspects is the reliable estimation of increasing and decreasing parts in the course of F0. This aspect is related to the central question: *When is a F0 movement perceptually important and when not?* Before dealing in detail with this question, it is important to analyze the problems during the estimation of increases and decreases in the course of F0.

The following aspects have to be taken into consideration:

1. there are several ways to estimate increasing and decreasing parts in the course of F0, in dependence of the chosen method the estimation is more or less reliable,
2. segmental influences on the F0 track may disrupt the estimation, and
3. faulty F0 values may disrupt the estimation as well.

With respect to the first aspect the crucial point is the differentiation of perceptually important from perceptually not important deviations from a general trendline in the course of F0. As a result the following two parameters are important for the estimation of the status of an increase or a decrease:

1. duration of increase or decrease, and
2. amount of increase or decrease.

One simple method for estimating F0 increases would be the comparison of neighboring F0 values like:

$F0_{t_0} \geq F0_{t_{0-1}}$ where $F0_{t_0}$ represents a F0 value at point t_0 and $F0_{t_{0-1}}$ is the F0 value one frame before.

The duration of an increase could be calculated by summing up the number of successively fulfilled cases. Whenever the criterion is not fulfilled, the algorithm stops counting. However, when looking at real F0 tracks it becomes apparent that this criterion is incomplete, since it is often the case that slight deviations from this criterion can be observed, but nevertheless a clear increase is visible. The latter statement is important, because often serves the visible inspection of the F0 track (cf. page 60) as decision criteria whether there is an increase during manual labeling. However, unlike the human ability to smooth certain deviations from a general trendline, the algorithm does not have this ability and would simply stop counting whenever the criterion is not fulfilled. Other methods have to be used as to account

for this aspect. Therefore, the approach presented includes an algorithm that allows a limited number of outlying values within a continuous part of F0 values. There are of course other mathematical methods of estimating increases and decreases in a times series of numerical values, for instance fitting a straight line to a number of values, etc. However, it became obvious during the development of the increase and decrease estimation algorithm that more complex criteria does often not provide the selectivity and the transparency of more simpler methods and therefore no attempt was made to integrate more complex algorithms for this aspect.

Second, the estimation of the size of an increase or decrease is usually bound to the problem of estimating the duration of increasing or decreasing parts. Whenever the estimation of the duration of increasing or decreasing stretches is inadequate, the estimation of the size of the change (decrease or increase) will be as well. One of the methods of estimating the size of a change before a maximum is the comparison of the maximum value with a preceding value that is at the beginning of the increase. The ratio of these two values could then represent the size of increase before the maximum. However, as noted before the adequate determination of the beginning of the increase is necessary for this method and runs into the same problems as during the estimation of the duration of an increase or decrease in F0. Therefore the size estimation algorithm uses the output of the algorithm described above that allows a certain number of outlying values in a continuous part of F0 values as basis for the calculation of ratios between the maximum or minimum and the beginning or end of an increase or decrease.

As a consequence of the considerations mentioned above the following parameters in the F0 domain are extracted by the program. The individual parameters are presented in detail, for an overview see table 5.2:

Number of continuously increasing or decreasing F0 values before or after point t_0:

Parameter	Range	Name
nr of continuously increasing F0 before t_0	0-20	*F0inB*
nr of continuously decreasing F0 before t_0	0-20	*F0deB*
nr of continuously decreasing F0 after t_0	0-20	*F0inA*
nr of continuously increasing F0 after t_0	0-20	*F0deA*

These parameters measure the number of continuously increasing or decreasing F0 values before or after point t_0 with the formulas

- F0inB: $F0_{t_0} \geq F0_{t_{0-1}}$ (increase to point t_0 before) or
- F0deB: $F0_{t_0} \leq F0_{t_{0-1}}$ (decrease to point t_0 before) or

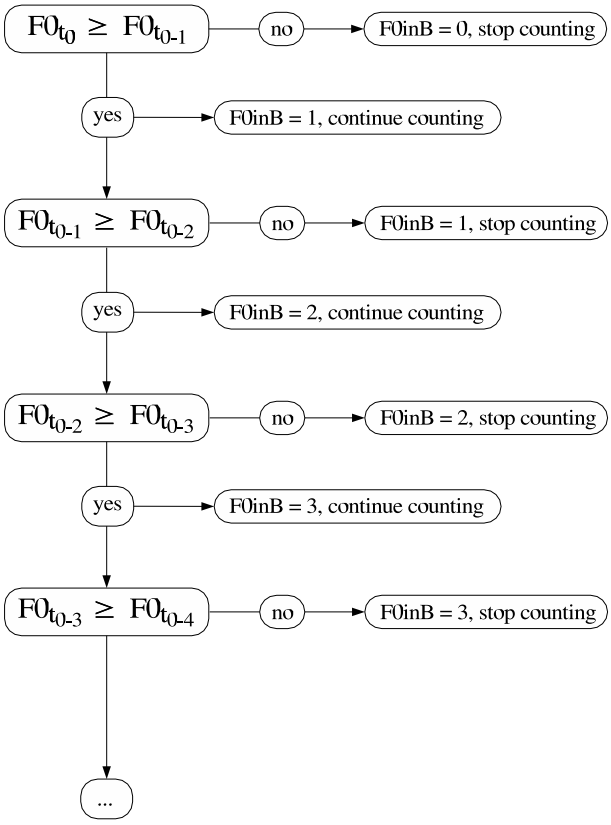


Figure 5.4: Depiction of the method of F0 increase estimation. To estimate an increase before frame $t_0$neighboring values preceding it are compared. If $F0_{t_0} \geq F0_{t_{0-1}}$ is fulfilled F0inB is set to one and the next comparsion is made. If $F0_{t_0} < F0_{t_{0-1}}$ then F0inB is set to zero and counting stops. Estimation of decreases works analogous.

- F0inA: $F0_{t_0} \leq F0_{t_{0+1}}$ (increase after point t_0) or
- F0deA: $F0_{t_0} \geq F0_{t_{0+1}}$ (decrease after point t_0).

The formulas represent solely the initial comparisons, which are continued successively until the condition is not fulfilled or the end of the detection interval has been reached (see figure 5.4) . Values may range from 0-20 frames since increases or decreases in F0 usually do not exceed this time domain. These parameters locate continuous increases or decreases in F0 but have the disadvantage that they might be stopped by just one F0 value out of sequence. To account for cases with outlying values another parameter (*F0in2B* see below) was introduced that allows a certain number of outlying values in a certain domain. Since the outputs of both of the increase or decrease estimation algorithms (without and with outlying values) are of interest here both where used.

Amount of F0 increase or decrease before or after point t_0:

Parameter	Range	Name
amount of F0 increase before t_0	1-$\star$	*AF0inB*
amount of F0 decrease after t_0	1-$\star$	*AF0deA*
amount of F0 decrease before t_0	0-1	*AF0deB*
amount of F0 increase after t_0	0-1	*AF0inA*

These parameters calculate the amount of increase or decrease before or after point t_0 by the ratio of the F0 value at point t_0 and the F0 value at the start or end of the increase or decrease before. As formulas

- $AF0inB = \frac{F0_{t_0}}{F0_{startIncr}}$ in which $F0_{startIncr} = F0_{t_0 - F0inB}$ and
- $AF0deB = \frac{F0_{t_0}}{F0_{startDecr}}$ in which $F0_{startDecr} = F0_{t_0 - F0deB}$ and
- $AF0inA = \frac{F0_{t_0}}{F0_{endIncr}}$ in which $F0_{endIncr} = F0_{t_0 + F0inA}$ and
- $AF0deA = \frac{F0_{t_0}}{F0_{endDecr}}$ in which $F0_{endDecr} = F0_{t_0 + F0deA}$.

Values may range from $1<AF0inB$ or $1<AF0deA$ and $0<AF0deB<1$ and $0<AF0inA<1$. See figure 5.5 for an illustration of this method. The relative comparison avoids problems that could appear when setting absolute comparison values.

Number of increasing or decreasing F0 values before or after point t_0 allowing outlying values:

Parameter	Range	Name
nr of increasing F0 before t_0 (allow outlying values)	0-20	*F0inB2*
nr of decreasing F0 before t_0 (allow outlying values)	0-20	*F0de*B2
nr of decreasing F0 after t_0 (allow outlying values)	0-20	*F0de*A2
nr of increasing F0 after t_0 (allow outlying values)	0-20	*F0inA2*

These parameters do also estimate the number of increasing or decreasing F0 values before or after point t_0 as the parameters already presented one section above. However, they do not require continuous increases or decreases, but allow outlying values. The algorithm allows up to 3 outlying values in a row before or after point t_0. Outlying values are cases that do not fulfill the formulas presented one section above, for instance where the formula $F0_{t_0} \geq F0_{t_{0-1}}$ (for the case of an increase to point t_0 before) is not fulfilled. Values may range from 0-20 frames.

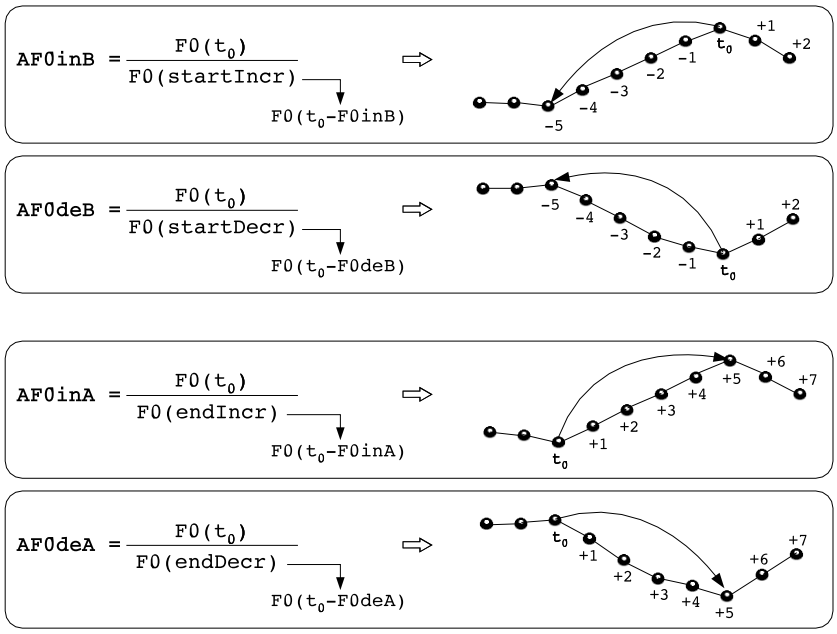


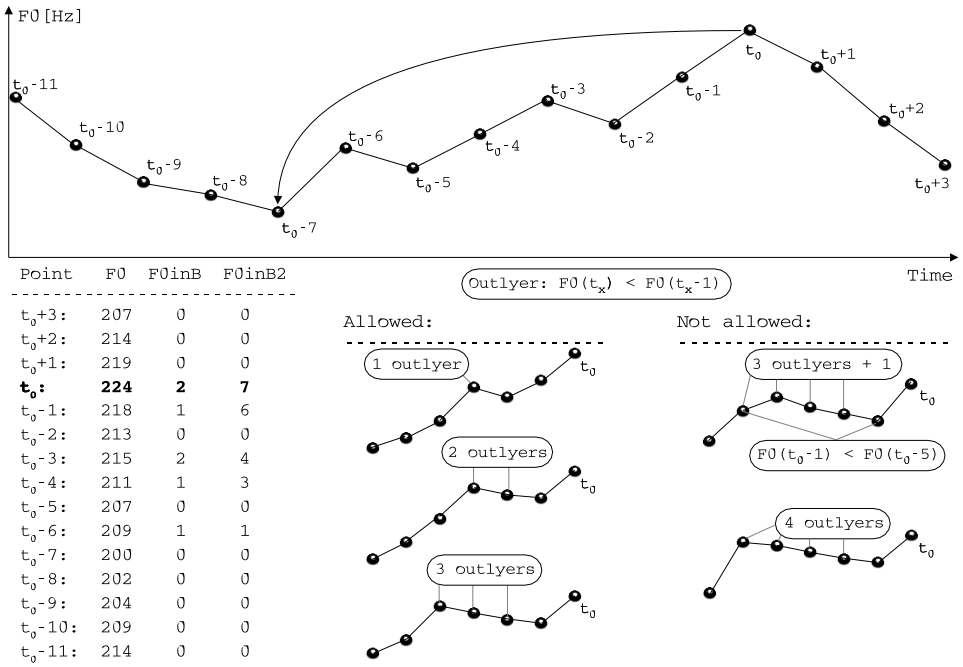
Figure 5.5: Illustration of the estimation method of the amounts of increases or decreases in the F0 track before (top two boxes) or after (bottom two boxes) point t_0.

These parameters are intended to give a better representation of increasing or decreasing parts before or after a F0 value under inspection since they allow outlying values from a strict increase or decrease. Figure 5.6 shows an illustration of the working method of this algorithm and compares it with the first estimation method. The decision to take 3 outlying values as allowable deviation is based on a number of cases that have been estimated by the algorithm and simultaneously checked by visual inspection. Sometimes one could allow more outlying values to categorise certain F0 movements correctly, but in order to avoid misinterpretations it was decided to restrict the number of outlying values to 3.

Amount of F0 increase or decrease before or after point t_0 allowing outlying values:

Parameter	Range	Name
amount of F0 increase before t_0 (allow outlying values)	1-$\star$	*AF0inB2*
amount of F0 decrease after t_0 (allow outlying values)	1-$\star$	*AF0deA2*
amount of F0 decrease before t_0 (allow outlying values)	0-1	*AF0deB2*
amount of F0 increase after t_0 (allow outlying values)	0-1	*AF0inA2*

These parameters calculate the amount of increase or decrease before or after point



Point	F0	F0inB	F0inB2
t_0+3:	207	0	0
t_0+2:	214	0	0
t_0+1:	219	0	0
t_0:	**224**	**2**	**7**
t_0-1:	218	1	6
t_0-2:	213	0	0
t_0-3:	215	2	4
t_0-4:	211	1	3
t_0-5:	207	0	0
t_0-6:	209	1	1
t_0-7:	200	0	0
t_0-8:	202	0	0
t_0-9:	204	0	0
t_0-10:	209	0	0
t_0-11:	214	0	0

Figure 5.6: Comparison of increase estimation algorithms F0inB vs. F0inB2. The same principle is used for the estimation of the other parameters F0deB2, F0inA2, and F0deA2. The table below the F0 graph shows the absolute F0 values and compares the output of F0inB and F0inB2. In the presented case F0inB (that does not allow outlying values from the criteria $F0_{t_x} \geq F0_{t_{x-1}}$) calculates only an increase of 2 frames before t_0 whereas F0inB2 calculates an increase of 7 frames before since the latter algorithm allows up to 3 successive outlying values. The graphs right of the table illustrate the number of allowed outlying values as well as two cases where the algorithm stops counting.

t_0 by the ratio of the F0 value at point t_0 and the F0 value at the start or end of the increase or decrease before as estimated by the method outlined above, namely by allowing a certain number of outlying values in a certain domain. As formulas

- $AF0inB2 = \frac{F0_{t_0}}{F0_{startIncr}}$ in which $F0_{startInc} = F0_{t_0 - F0inB2}$ and
- $AF0deB2 = \frac{F0_{t_0}}{F0_{startDecr}}$ in which $F0_{startDecr} = F0_{t_0 - F0deB2}$ and
- $AF0inA2 = \frac{F0_{t_0}}{F0_{endIncr}}$ in which $F0_{endIncr} = F0_{t_0 + F0inA2}$ and
- $AF0deA2 = \frac{F0_{t_0}}{F0_{endDecr}}$ in which $F0_{endDecr} = F0_{t_0 + F0deA2}$.

Values may range from $1 < AF0inB2$ or $1 < AF0deA2$ and $0 < AF0deB2 < 1$ and $0 < AF0inA2 < 1$.

Number of smaller F0 values before or after point t_0 within an interval:

Parameter	Range	Name
nr of smaller F0 before t_0, 5, no voicing control	0-5	*F0sno5B*
nr of smaller F0 before t_0, 10, no voicing control	0-10	*F0sno10B*
nr of smaller F0 before t_0, 16, no voicing control	0-16	*F0sno16B*
nr of smaller F0 before t_0, 23, no voicing control	0-23	*F0sno23B*
nr of smaller F0 before t_0, 31, no voicing control	0-31	*F0sno31B*
nr of smaller F0 before t_0, 40, no voicing control	0-40	*F0sno40B*
nr of smaller F0 after t_0, 5, no voicing control	0-5	*F0sno5A*
nr of smaller F0 after t_0, 10, no voicing control	0-10	*F0sno10A*
nr of smaller F0 after t_0, 16, no voicing control	0-16	*F0sno16A*
nr of smaller F0 after t_0, 23, no voicing control	0-23	*F0sno23A*
nr of smaller F0 after t_0, 31, no voicing control	0-31	*F0sno31A*
nr of smaller F0 after t_0, 40, no voicing control	0-40	*F0sno40A*

These parameters calculate the number of smaller F0 values before or after point t_0 (see figure 5.7) within predefined intervals without voicing control, that is, it is not differentiated between F0 values equal 0 (= unvoiced values) and F0 values > 0 (= voiced). Expressed in formulas:

- F0snoB: $F0_{t_0} > F0_{t_{0-x}}$ where x = 1-5, 1-10, 1-16, 1-23, 1-31, 1-40
- F0snoA: $F0_{t_0} > F0_{t_{0+x}}$ where x = 1-5, 1-10, 1-16, 1-23, 1-31, 1-40

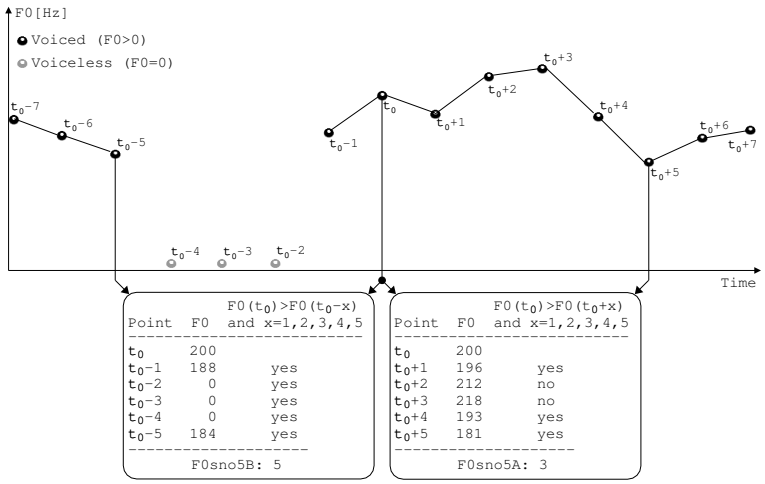


Figure 5.7: Illustration of the estimation method of the number of smaller F0 values before or after point t_0 in the F0 track without voicing control. The picture shows the estimation for the first interval size of 5 frames, that is F0sno5B for the number of smaller F0 values 5 frames before t_0 and F0sno5A for the ones after t_0 respectively.

The parameters are intended to give an estimation of the relative height of a given F0 value under inspection. The sizes of the intervals are the same as already mentioned for the estimation of the number of voiced values before and after point t_0. The interval sizes were chosen because they cover the whole 400 ms interval in 6 steps with successively increasing interval size from the smallest one with 5 frames up to the longest one with 40 frames, (5, 10, 16, 23, 31, 40, i.e. +5, +6, +7, +8, +9 respectively).

Number of smaller F0 values before or after point t_0 with voicing control:

Parameter	Range	Name
nr of smaller F0 before t_0, voicing control	0-40	*F0smB*
nr of smaller F0 after t_0, voicing control	0-40	*F0smA*

These parameters calculate the number of smaller F0 values before or after point t_0 within an uninterrupted voiced stretch, that is only F0 values that are not equal 0 are counted. Therefore, these parameters give exact numbers of smaller F0 values before or after point t_0 within a voiced phase.

- F0smB: $F0_{t_0} \geq F0_{t_{0-x}}$ where x may range from $1-40$ as long as $F0_{t_{0-x}} > 0$
- F0smA: $F0_{t_0} \geq F0_{t_{0+x}}$ where x may range from $1-40$ as long as $F0_{t_{0+x}} > 0$

Values may range from 0-40 frames. Here the comparison is not made between neighboring values but always from point t_0 to successively preceding or following values and does not stop when there is a larger value but simply goes on until the end of the continuously voiced part.

Number of F0 values before or after point t_0 within ± 7 Hz:

Parameter	Range	Name
nr of F0 before t_0 within ± 7 Hz	0-40	*F0intB*
nr of F0 after t_0 within ± 7 Hz	0-40	*F0intA*

These parameters calculate the number of F0 values before or after point t_0 that lie within a ± 7 Hz interval. The summation stops when a value falls outside the bounds of the interval. Values may range from 0-40 frames. The frequency range of ± 7 Hz was chosen as a result of an estimation of the range of F0 values that appear as smooth during visual inspection. Therefore, this value is intended to support the decision whether the course of F0 before or after is smooth or rapid.

5.3.3 Parameters in the RMS Domain

A number of additional parameters in the RMS domain have been introduced basically in order to account for the improved estimation of boundary tones. The parameters in the RMS domain are used to aquire additional cues for the decision whether a given F0 movement may be considered as pitch accent or boundary tone candidate or not (cf. figure 4.5). Furthermore, the RMS features play a crucial role in the location of intonation phrase boundaries. The parameters will be presented next in detail, for an overview of them see table 5.3)

Number of increasing or decreasing RMS values before or after point t_0:

Parameter	Range	Name
nr of increasing RMS values before t_0	0-20	*RMinB*
nr of decreasing RMS values before t_0	0-20	*RMdeB*
nr of increasing RMS values after t_0	0-20	*RMinA*
nr of decreasing RMS values after t_0	0-20	*RMdeA*

The estimation of increases or decreases in RMS amplitude usually does not pose such problems as in the case of F0. However, sometimes there can also be single

outlying values and to account for these occasions the number of allowed outlying values was restricted to maximally 1. The number of increasing or decreasing RMS values before or after point t_0 is calculated by these parameters. One outlier is allowed. Values may range from 0-20 frames.

- $RMinB = RMS_{t_0} > RMS_{t_0-1}$ and $RMdeB = RMS_{t_0} < RMS_{t_0-1}$
- $RMinA = RMS_{t_0} < RMS_{t_0+1}$ and $RMdeA = RMS_{t_0} > RMS_{t_0+1}$

Amount of increasing or decreasing RMS values before or after point t_0:

Parameter	Range	Name
amount of increase in RMS before t_0	1-$\star$	*ARMinB*
amount of decrease in RMS after t_0	1-$\star$	*ARMdeA*
amount of decrease in RMS before t_0	0-1	*ARMdeB*
amount of increase in RMS after t_0	0-1	*ARMinA*

The amount of increasing or decreasing RMS values before or after point t_0 is calculated by the ratio of the RMS value at point t_0 and the beginning of the increase or decrease before or the end of the increase or decrease after it. As formulas:

- $ARMinB = \frac{RMS_{t_0}}{RMS_{startIncr}}$ in which $RMS_{startIncr} = RMS_{t_0-RMinB}$
- $ARMdeB = \frac{RMS_{t_0}}{RMS_{startDecr}}$ in which $RMS_{startDecr} = RMS_{t_0-RMdeB}$
- $ARMinA = \frac{RMS_{t_0}}{RMS_{endIncr}}$ in which $RMS_{endIncr} = RMS_{t_0-RMinA}$
- $ARMdeA = \frac{RMS_{t_0}}{RMS_{endDecr}}$ in which $RMS_{endDecr} = RMS_{t_0-RMdeA}$

Values may range from $1 < ARMinB$ or $1 < ARMdeA$ and $0 < ARMdeB < 1$ and $0 < ARMinA < 1$.

Number of smaller RMS values before or after point t_0 within an interval:

These parameters calculate the number of smaller RMS values before or after point t_0 within predefined intervals. The parameters are intended to give an estimation of the relative height of a given RMS value under inspection. The algorithm simply counts the number of smaller values and does not stop counting when there is a larger one.

Since the parameters in the first parameter assessment experiment were inadequate for the boundary tone detection the following 3 parameters where added.

Parameter	Range	Name
nr of smaller RMS values 5 before t_0	0-5	*RMsm5B*
nr of smaller RMS values 10 before t_0	0-10	*RMsm10B*
nr of smaller RMS values 16 before t_0	0-16	*RMsm16B*
nr of smaller RMS values 23 before t_0	0-23	*RMsm23B*
nr of smaller RMS values 31 before t_0	0-31	*RMsm31B*
nr of smaller RMS values 40 before t_0	0-40	*RMsm40B*
nr of smaller RMS values 5 after t_0	0-5	*RMsm5A*
nr of smaller RMS values 10 after t_0	0-10	*RMsm10A*
nr of smaller RMS values 16 after t_0	0-16	*RMsm16A*
nr of smaller RMS values 23 after t_0	0-23	*RMsm23A*
nr of smaller RMS values 31 after t_0	0-31	*RMsm31A*
nr of smaller RMS values 40 after t_0	0-40	*RMsm40A*

Number of RMS values before or after point t_0 with small changes:

Parameter	Range	Name
nr of RMS values before t_0 with small changes	0-40	*RMscB*
nr of RMS values after t_0 with small changes	0-40	*RMscA*

These parameters calculate the number of RMS values before or after point t_0 that do not change much, that is, lie within an interval of $\pm$128 RMS amplitude and are $<$500 RMS amplitude. The exact numbers were established by the inspection of a large number of intonation phrase boundaries. The fact that the comparisons are made between RMS value at point t_0 and all the following or preceding values is of importance.

The parameters are intended to give an estimation of a part with very small amplitude changes like pauses as compared to parts including speech that usually includes larger amplitude movements. The greater the number, the longer there is no strong change in the RMS values which might indicate a pause for instance.

Number of RMS values before or after point t_x that are a certain percentage smaller than the RMS value at point t_x:

These parameters calculate the number of RMS values before or after a point t_x that are a certain percentage smaller than the RMS value at point t_x. Point t_x is located as follows: when the percentage of smaller values before is calculated, then point

Parameter	Range	Name
nr of a certain percentage smaller RMS values before t_x	0-40	*RMpsB*
nr of a certain percentage smaller RMS values after t_x	0-40	*RMpsA*

t_x is point $t_0 + RMinA$, the next maximum at the end of an increase after point t_0. When the percentage of smaller values afterwards is calculated, then point t_x is the point $t_0 + RMinB$, the last maximum in RMS before point t_0. Expressed in formulas:

- $RMpsB = RMS_{t_0+RMinA} * 0.7 > RMS_{t_0}$
- $RMpsA = RMS_{t_0+RMinB} * 0.7 > RMS_{t_0}$

The percentage of smaller RMS values (*RMpsB*) before is established by multiplying the RMS value at the end of an increase by 0.7 and comparing it with the actual RMS values. That is, a maximum in the course of RMS is compared with its preceding values. The exact numbers were established by the inspection of a large number of intonation phrase boundaries.

Number of RMS values before or after point t_0 within a predefined range:

Parameter	Range	Name
nr of RMS values before t_0 within a predefined range	0-30	*RMsmrB*
nr of RMS values after t_0 within a predefined range	0-30	*RMsmrA*

These parameters calculate the number of RMS values before or after point t_0 that fall within a predefined range of ± 128 RMS amplitude. Expressed in formulas:

- $RMsmrB = RMS_{t_0} - RMS_{t_0-1} > -128 \;\&\; RMS_{t_0} - RMS_{t_0-1} < 128$
- $RMsmrA = RMS_{t_0} - RMS_{t_0+1} > -128 \;\&\; RMS_{t_0} - RMS_{t_0+1} < 128$

The parameters are intended to represent positions of possible intonation phrase breaks.

Voicing		
Parameter	Range	Name
nr of continuously voiced values before t_0	0-40	*VoicB*
nr of continuously voicel. values before t_0	0-40	*VoilB*
nr of continuously voiced values after t_0	0-40	*VoicA*
nr of continuously voicel. values after t_0	0-40	*VoilA*
nr of voiced values 50 ms before t_0	0-5	*Voic5B*
nr of voiced values 100 ms before t_0	0-10	*Voic10B*
nr of voiced values 160 ms before t_0	0-16	*Voic16B*
nr of voiced values 230 ms before t_0	0-23	*Voic23B*
nr of voiced values 310 ms before t_0	0-31	*Voic31B*
nr of voiced values 400 ms before t_0	0-40	*Voic40B*
nr of voiced values 50 ms after t_0	0-5	*Voil5A*
nr of voiced values 100 ms after t_0	0-10	*Voil10A*
nr of voiced values 160 ms after t_0	0-16	*Voil16A*
nr of voiced values 230 ms after t_0	0-23	*Voil23A*
nr of voiced values 310 ms after t_0	0-31	*Voil31A*
nr of voiced values 400 ms after t_0	0-40	*Voil40A*

Table 5.1: Table of parameters in the voicing domain.

5.3.4 Summary

All criteria were implemented in a computer program that extracted the individual values for each of the pitch accents and boundary tones from the GToBI corpus automatically. The output could be directly imported into a statistics program for further processing. In the following section the results of the parameter assessment program are presented and discussed.

5.3.5 Results

The results of the parameter assessment program were used for statistical processing in order to formulate adequate detection criteria for the acoustic features. From each of the acoustic features the mean, median, standard deviation, as well as the minimal and maximal values were calculated with a standard statistics program (StarOffice 5.2 Calc). Altogether 74 parameters were extracted, 16 in the voicing domain, 32 in the F0 domain, and 26 in the RMS domain.

F0		
Parameter	Range	Name
nr of continuously increasing F0 before t_0	0-20	*F0inB*
nr of continuously decreasing F0 before t_0	0-20	*F0deB*
nr of continuously decreasing F0 after t_0	0-20	*F0inA*
nr of continuously increasing F0 after t_0	0-20	*F0deA*
amount of F0 increase before t_0	1-$\star$	*AF0inB*
amount of F0 decrease after t_0	1-$\star$	*AF0deA*
amount of F0 decrease before t_0	0-1	*AF0deB*
amount of F0 increase after t_0	0-1	*AF0inA*
nr of increasing F0 before t_0 (allow outlying values)	0-20	*F0inB2*
nr of decreasing F0 before t_0 (allow outlying values)	0-20	*F0deB2*
nr of decreasing F0 after t_0 (allow outlying values)	0-20	*F0deA2*
nr of increasing F0 after t_0 (allow outlying values)	0-20	*F0inA2*
amount of F0 increase before t_0 (allow outlying values)	1-$\star$	*AF0inB2*
amount of F0 decrease after t_0 (allow outlying values)	1-$\star$	*AF0deA2*
amount of F0 decrease before t_0 (allow outlying values)	0-1	*AF0deB2*
amount of F0 increase after t_0 (allow outlying values)	0-1	*AF0inA2*
nr of smaller F0 before t_0, 5, no voicing control	0-5	*F0sno5B*
nr of smaller F0 before t_0 , 10, no voicing control	0-10	*F0sno10B*
nr of smaller F0 before t_0, 16, no voicing control	0-16	*F0sno16B*
nr of smaller F0 before t_0, 23, no voicing control	0-23	*F0sno23B*
nr of smaller F0 before t_0, 31, no voicing control	0-31	*F0sno31B*
nr of smaller F0 before t_0, 40, no voicing control	0-40	*F0sno40B*
nr of smaller F0 after t_0, 5, no voicing control	0-5	*F0sno5A*
nr of smaller F0 after t_0, 10, no voicing control	0-10	*F0sno10A*
nr of smaller F0 after t_0, 16, no voicing control	0-16	*F0sno16A*
nr of smaller F0 after t_0, 23, no voicing control	0-23	*F0sno23A*
nr of smaller F0 after t_0, 31, no voicing control	0-31	*F0sno31A*
nr of smaller F0 after t_0, 40, no voicing control	0-40	*F0sno40A*
nr of smaller F0 before t_0, voicing control	0-40	*F0smB*
nr of smaller F0 after t_0, voicing control	0-40	*F0smA*
nr of F0 before t_0 within $\pm$7 Hz	0-40	*F0intB*
nr of F0 after t_0 within $\pm$7 Hz	0-40	*F0intA*

Table 5.2: Table of parameters in the F0 domain.

RMS		
Parameter	Range	Name
nr of increasing RMS values before t_0	0-20	*RMinB*
nr of decreasing RMS values before t_0	0-20	*RMdeB*
nr of increasing RMS values after t_0	0-20	*RMinA*
nr of decreasing RMS values after t_0	0-20	*RMdeA*
amount of increase in RMS before t_0	1-$\star$	*ARMinB*
amount of decrease in RMS after t_0	1-$\star$	*ARMdeA*
amount of decrease in RMS before t_0	0-1	*ARMdeB*
amount of increase in RMS after t_0	0-1	*ARMinA*
nr of smaller RMS values 5 before t_0	0-5	*RMsm5B*
nr of smaller RMS values 10 before t_0	0-10	*RMsm10B*
nr of smaller RMS values 16 before t_0	0-16	*RMsm16B*
nr of smaller RMS values 23 before t_0	0-23	*RMsm23B*
nr of smaller RMS values 31 before t_0	0-31	*RMsm31B*
nr of smaller RMS values 40 before t_0	0-40	*RMsm40B*
nr of smaller RMS values 5 after t_0	0-5	*RMsm5A*
nr of smaller RMS values 10 after t_0	0-10	*RMsm10A*
nr of smaller RMS values 16 after t_0	0-16	*RMsm16A*
nr of smaller RMS values 23 after t_0	0-23	*RMsm23A*
nr of smaller RMS values 31 after t_0	0-31	*RMsm31A*
nr of smaller RMS values 40 after t_0	0-40	*RMsm40A*
nr of RMS values before t_0 with small changes	0-40	*RMscB*
nr of RMS values after t_0 with small changes	0-40	*RMscA*
nr of a certain percentage smaller RMS values before t_x	0-40	*RMpsB*
nr of a certain percentage smaller RMS values after t_x	0-40	*RMpsA*
nr of RMS values before t_0 within a predefined range	0-30	*RMsmrB*
nr of RMS values after t_0 within a predefined range	0-30	*RMsmrA*

Table 5.3: Table of parameters in the RMS domain. See page 119 for the explanation of t_x.

A summary of the results for all the pitch accents in the GToBI training material is presented in tables 5.4 (voicing), 5.5 (F0), 5.6 (RMS) and for all the boundary tones in tables 5.7 (voicing), 5.8 (F0), and 5.9 (RMS). The results will now be discussed in more detail first for the pitch accents and second for the boundary tones.

Results for pitch accents

As in the first parameter analysis the large variability in most of the parameters is expressed by large standard deviations. Since some of the results have already been discussed in section 4.3 the focus will be on the new parameters. The differences between the two algorithms estimating the duration of increases or decreases are visible in the following cases (see table 5.5): the numbers given represent frames. That is, L+H* (F0inB: 8 vs. F0inB2: 11) reads as the median increase in F0 for L+H* pitch accents is 8 frames from the beginning of the increase up to the maximum in the case of the method applied for F0inB versus 11 frames in the case of the second method F0inB2 allowing outlying values. Since the frames are given every 10 ms the number may be multiplied by 10 and the result represents milliseconds. Also the H+!H* (F0deA: 1 vs. F0deA2: 5) and L* (F0inA: 9 vs. F0inA2: 14) pitch accents show large differences between the two estimation methods. As a consequence the values of the amount estimations are also different and a visual inspection of the corresponding F0 tracks verified that the second estimation algorithm that allowed a limited number of outlying values delivers more representative results. L* pitch accents seem to be characterized by a fairly steep F0 increase afterwards (F0inA2: 14 and AF0inA2: 0.51, where the last number represents the median ratio of F0 values: F0 at the beginning of the increase divided by F0 at the end of the increase).

As one would expect, the number of F0 values that are smaller before and after than the actual F0 value at the pitch accent is high for the high pitch accents (e.g., H*: F0sno40B: 35, F0sno40A: 35) and small for the low pitch accents (e.g., L*: F0sno40B: 24, F0sno40A: 15). Exceptions from the latter rule are the H+!H* that has more larger F0 values before (as one would expect as a result of downstep, F0sno40B: 14) and the H+L* that has about the same number of smaller F0 values afterwards than the high pitch accents. The last result can be explained when checking the corresponding F0 tracks from the files that include the six H+L* cases: all of them are marked before a final L-L% boundary tone and are most often in a falling F0 part.

As expected, the median number of smaller F0 values before with voicing control is highest for the L+H* cases (F0smB: 14). However, it is unexpectedly the same (6) for the H* and L* accents. This indicates once more that low pitch accents are not always labeled at F0 minima.

The results in the RMS domain (see table 5.6) with respect to the RMS increase or decrease estimation as well as the amount estimation did not change significantly as

VOICING												
Tone	**H***		**L+H***		**H+!H***		**L***		**L*+H**		**H+L***	
Nr items	51		25		7		11		7		6	
	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD
VoicB	7	(13)	15	(12)	11	(9)	17	(13)	9	(8)	23	(12)
VoilB	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
VoicA	13	(13)	19	(13)	40	(14)	24	(12)	10	(15)	10	(15)
VoilA	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
Voic5B	5	(.9)	5	(.2)	5	(0)	5	(.7)	5	(0)	5	(1)
Voic10B	9	(3)	10	(2)	10	(2)	10	(3)	9	(2)	10	(4)
Voic16B	9	(5)	15	(4)	11	(4)	16	(5)	9	(4)	16	(6)
Voic23B	15	(7)	20	(5)	15	(5)	17	(7)	15	(5)	20	(7)
Voic31B	20	(8)	23	(7)	19	(5)	17	(9)	18	(7)	22	(7)
Voic40B	27	(10)	28	(9)	26	(5)	17	(12)	25	(10)	29	(7)
Voic5A	5	(1)	5	(1)	5	(0)	5	(0)	5	(1)	5	(0)
Voic10A	10	(3)	10	(3)	10	(2)	10	(1)	10	(2)	10	(2)
Voic16A	13	(5)	16	(5)	16	(3)	16	(3)	12	(3)	14	(4)
Voic23A	17	(7)	19	(7)	23	(5)	23	(6)	16	(4)	18	(6)
Voic31A	25	(7)	26	(8)	31	(7)	24	(8)	22	(5)	22	(8)
Voic40A	29	(8)	31	(9)	40	(10)	26	(11)	30	(6)	30	(11)

Table 5.4: Median values (Md) and standard deviations (SD, in brackets) for the acoustic features in the **voicing** domain for the pitch accents in the GToBI training material.

a result of the allowance of one outlying value in increases or decreases. Therefore the discussion focuses on the new parameters.

The number of smaller RMS values before and after indicate that pitch accents are usually marked at positions that are prominent in the sense of high RMS values or in other words: pitch accents are usually associated with energy maxima in their immediate vicinity. The values for the number of RMS values with small changes and the percentage of smaller values are all 0 as one would expect because these parameters should only count at the beginning and end of intonation phrases or speech pauses. The values of RMsmrB/A are also very small and mostly do not exceed 1, except in the H+L* cases it seems that here are small RMS movements before (RMsmrB: 5) and after (RMsmrA: 7).

Results for boundary tones

The results of the parameter assessment program for the boundary tones in the GToBI corpus are summarized in tables 5.7, 5.8, and 5.9. First, the four intona-

F0												
Tone	**H***		**L+H***		**H+!H***		**L***		**L*+H**		**H+L***	
Nr items	51		25		7		11		7		6	
	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD
F0inB	3	(4)	8	(4)	0	(1)	1	(3)	0	(3)	0	(.5)
F0deB	0	(1)	0	(0)	1	(4)	0	(3)	1	(5)	3	(2)
F0deA	0	(4)	0	(5)	1	(3)	0	(1)	1	(1)	3	(2)
F0inA	1	(3)	1	(2)	0	(1)	9	(8)	0	(2)	0	(0)
AF0inB	1.04	(.5)	1.17	(.1)	0	(.5)	1.01	(.5)	0	(.6)	0	(.5)
AF0deA	0	(.6)	0	(.6)	1.01	(.7)	0	(.5)	1	(.5)	1.09	(.5)
AF0deB	0	(.4)	0	(0)	0.75	(.4)	0	(.4)	0.79	(.5)	0.81	(.4)
AF0inA	0.93	(.5)	0.77	(.5)	0	(.4)	0.51	(.3)	0	(.5)	0	(0)
F0inB2	3	(4)	11	(5)	0	(2)	2	(4)	0	(2)	0	(.4)
F0deB2	0	(1)	0	(0)	1	(4)	0	(3)	2	(7)	3	(8)
F0deA2	0	(6)	0	(6)	5	(7)	0	(3)	1	(1)	5	(2)
F0inA2	1	(3)	1	(4)	0	(2)	14	(8)	0	(5)	0	(2)
AF0inB2	1.05	(.5)	1.28	(.3)	0	(.5)	1.02	(.6)	0	(.5)	0	(.4)
AF0deA2	0	(.6)	0	(1)	1.1	(.7)	0	(.5)	1	(.5)	1.07	(.5)
AF0deB2	0	(.4)	0	(0)	0.75	(.4)	0	(.4)	0.64	(.4)	0.76	(.3)
AF0inA2	0.91	(.5)	0.77	(.4)	0	(.4)	0.51	(.3)	0	(.4)	0	(.3)
F0sno5B	5	(1)	5	(.4)	2	(2)	3	(2)	2	(2)	2	(1)
F0sno10B	9	(2)	10	(.6)	3	(3)	8	(3)	4	(4)	2	(3)
F0sno16B	15	(3)	16	(2)	8	(5)	12	(5)	10	(6)	2	(6)
F0sno23B	22	(5)	23	(2)	12	(7)	16	(5)	14	(8)	5	(6)
F0sno31B	28	(7)	31	(2)	12	(7)	19	(7)	19	(10)	10	(7)
F0sno40B	35	(10)	39	(4)	14	(6)	24	(9)	20	(10)	11	(6)
F0sno5A	3	(2)	4	(2)	4	(1)	0	(2)	1	(2)	5	(2)
F0sno10A	7	(3)	8	(4)	9	(3)	0	(4)	2	(3)	9	(2)
F0sno16A	13	(4)	14	(6)	15	(3)	0	(6)	5	(3)	15	(3)
F0sno23A	19	(6)	21	(8)	20	(3)	2	(8)	7	(4)	22	(5)
F0sno31A	27	(8)	29	(10)	28	(4)	9	(9)	10	(7)	26	(7)
F0sno40A	35	(8)	34	(10)	37	(6)	15	(11)	11	(9)	32	(10)
F0smB	6	(11)	14	(11)	2	(2)	6	(13)	2	(6)	0	(1)
F0smA	8	(12)	7	(14)	27	(15)	0	(6)	1	(2)	10	(11)
F0intB	5	(8)	4	(9)	6	(3)	9	(10)	6	(9)	3	(4)
F0intA	7	(11)	3	(7)	11	(12)	5	(10)	10	(15)	7	(11)

Table 5.5: Median values (Md) and standard deviations (SD, in brackets) for the acoustic features in the **F0** domain for the pitch accents in the GToBI training material.

RMS												
Tone	**H***		**L+H***		**H+!H***		**L***		**L*+H**		**H+L***	
Nr items	51		25		7		11		7		6	
	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD
RMinB	0	(4)	0	(3)	0	(3)	1	(7)	1	(5)	4	(1)
RMdeB	1	(2)	1	(2)	1	(3)	0	(2)	0	(1)	0	(0)
RMinA	0	(1)	0	(.3)	0	(.4)	1	(4)	0	(1)	1	(.5)
RMdeA	4	(4)	5	(4)	6	(4)	0	(2)	1	(3)	0	(3)
ARMinB	0	(66)	0	(6)	0	(8)	1.09	(8)	1.01	(8)	3.65	(4)
ARMdeA	3.04	(14)	3.37	(.13)	3.52	(58)	0	(5)	1.03	(13)	0	(2)
ARMdeB	0.52	(.4)	0.72	(.4)	0.73	(.5)	0	(.4)	0	(.4)	0	(9)
ARMinA	0	(.3)	0	(.3)	0	(.4)	0.4	(.4)	0	(.5)	0.92	(.4)
RMsm5B	2	(2)	2	(2)	4	(2)	5	(2)	3	(2)	5	(1)
RMsm10B	7	(3)	6	(4)	9	(3)	10	(3)	8	(3)	9	(3)
RMsm16B	13	(4)	11	(5)	15	(3)	16	(4)	14	(5)	13	(5)
RMsm23B	20	(4)	18	(5)	22	(3)	23	(4)	21	(6)	17	(5)
RMsm31B	27	(6)	26	(6)	30	(4)	31	(5)	29	(6)	19	(6)
RMsm40B	36	(7)	35	(7)	39	(5)	39	(5)	38	(6)	23	(6)
RMsm5A	5	(1)	5	(1)	5	(1)	2	(2)	4	(2)	3	(2)
RMsm10A	10	(2)	10	(1)	10	(1)	7	(4)	9	(4)	8	(3)
RMsm16A	16	(4)	16	(1)	16	(1)	11	(6)	15	(6)	14	(4)
RMsm23A	23	(5)	22	(2)	23	(1)	17	(7)	22	(8)	21	(6)
RMsm31A	29	(5)	29	(3)	31	(1)	25	(8)	30	(11)	29	(9)
RMsm40A	34	(6)	38	(5)	40	(1)	34	(11)	38	(13)	38	(11)
RMscB	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(1)
RMscA	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(.8)
RMpsB	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
RMpsA	0	(3)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
RMsmrB	1	(4)	1	(3)	1	(1)	1	(4)	1	(2)	5	(4)
RMsmrA	0	(3)	0	(6)	0	(1)	1	(9)	0	(3)	7	(10)

Table 5.6: Median values (Md) and standard deviations (SD, in brackets) for the acoustic features in the RMS domain for the pitch accents in the GToBI training material.

tion phrase boundary tones will be discussed and afterwards the two intermediate phrase tones. In the case of H-L% boundary tones there are only two items, which is not sufficient for representative statistics. However, they were listed for reasons of completeness. The discussion focuses on the new introduced parameters.

Intonation phrase boundary tones

One expects to get a larger number of voiced values (without continuation control) before the boundary since there is usually speech with its typical amplitude-variations and its successive changes of voiced and voiceless parts. This is reflected in the results with a large number of voiceless values after the boundaries (median values for VoilA: L-L%: 40; L-H%: 29, H-H%: 40, and H-L%: 25) and also in the parameters without continuation control (median values for Voic40B: L-L%: 33; L-H%: 20, H-H%: 35, H-L%: 21 and median values for Voic40A: L-L%: 0; L-H%: 11, H-H%: 0, H-L%: 13). Furthermore a large number of RMS values after the boundary location are smaller than the value there (median values for L-L%: 39; L-H%: 27; H-H%: 34; H-L%: 23).

As one would expect, the number of RMS values afterwards with small changes (RMscA) is much higher for the boundary tones than for the pitch accents (median values of RMscA: L-L%: 39; L-H%: 39; H-H%: 39; H-L%: 25). Also the number of RMS values that are a certain percentage smaller (RMSpsA) after point t_0 is higher for the boundary tones (median values: L-L%: 9; L-H%: 9; H-H%: 10; H-L%: 5) and as expected the number of RMS values within a small range afterwards (RMsmrA, median values: L-L%: 30; L-H%: 29; H-H%: 30; H-L%: 19).

Intermediate Phrase Boundary Tones

Here, the number of voiced items before without continuation control is similar to the intonation phrase boundary tones but the picture is different afterwards: the intermediate phrase boundary tones show much larger values (median values for Voic40A: L-: 26; H-: 31). Since the H- tones are sometimes labeled at the end of an F0 increase the number of smaller F0 values before without voicing control is fairly large (median value of F0sno40B for H-: 17).

In the cases of the RMS analysis the picture is not as clear cut as for the intonation phrase boundary tones. Especially the cases of RMscA and RMsmrA show much smaller values since the intermediate boundary tones are more often labeled at locations that do not have long pauses afterwards, as is often the case with intonation phrase-boundary tones.

To sum up, one can say that the results of the parameter acquisition program showed the advantages of the newly developed F0 increase or decrease estimation algorithm as well as the importance of the additional criteria in all three domains as compared to the initial analysis.

The results from the visual and auditory inspection of the corpus as well as the results from the automatic analysis of the underlying acoustic features of pitch

VOICING												
Tone	**L-L%**		**L-H%**		**H-H%**		**H-L%**		**L-**		**H-**	
nr items	27		5		8		2		10		25	
	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD
VoicB	0	(16)	0	(0)	18	(19)	0	(0)	0	(13)	24	(17)
VoilB	2	(7)	13	(3)	4	(8)	17	(3)	2	(12)	0	(6)
VoicA	0	(3)	0	(0)	0	(13)	0	(0)	0	(15)	4	(15)
VoilA	40	(18)	29	(11)	40	(16)	25	(16)	3	(12)	0	(8)
Voic5B	3	(2)	0	(0)	3	(3)	0	(0)	3	(2)	5	(2)
Voic10B	8	(4)	0	(0)	7	(5)	0	(0)	6	(4)	10	(4)
Voic16B	13	(5)	3	(3)	13	(7)	1	(1)	12	(5)	16	(6)
Voic23B	20	(7)	10	(3)	20	(8)	7	(3)	17	(7)	23	(7)
Voic31B	25	(7)	18	(3)	28	(8)	15	(3)	25	(10)	28	(7)
Voic40B	33	(9)	20	(1)	35	(8)	21	(1)	34	(12)	31	(9)
Voic5A	0	(2)	0	(0)	0	(2)	0	(0)	2	(2)	4	(2)
Voic10A	0	(3)	0	(0)	0	(3)	1	(.5)	6	(4)	6	(4)
Voic16A	0	(4)	0	(0)	0	(5)	4	(4)	8	(6)	10	(4)
Voic23A	0	(5)	3	(3)	0	(8)	7	(7)	15	(8)	15	(6)
Voic31A	0	(6)	7	(7)	0	(12)	11	(11)	21	(10)	23	(7)
Voic40A	0	(8)	11	(11)	0	(15)	13	(13)	26	(13)	31	(9)

Table 5.7: Median values (Md) and standard deviations (SD, in brackets) for the acoustic features in the **voicing** domain for the boundary tones in the GToBI training material.

F0												
Tone	**L-L%**		**L-H%**		**H-H%**		**H-L%**		**L-**		**H-**	
nr items	27		5		8		2		10		25	
	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD
F0inB	0	(.5)	0	(0)	0	(0)	0	(0)	0	(1)	0	(.8)
F0deB	0	(.8)	0	(0)	0	(2)	0	(0)	0	(3)	0	(2)
F0deA	0	(.8)	0	(0)	0	(0)	0	(0)	0	(2)	0	(1)
F0inA	0	(.3)	0	(0)	0	(.7)	0	(0)	0	(0)	0	(2)
AF0inB	0	(.3)	0	(0)	0	(0)	0	(0)	0	(.8)	0	(.5)
AF0deA	0	(.6)	0	(0)	0	(0)	0	(0)	0	(.6)	0	(.5)
AF0deB	0	(.4)	0	(0)	0	(.4)	0	(0)	0	(.2)	0	(.3)
AF0inA	0	(.3)	0	(0)	0	(.3)	0	(0)	0	(0)	0	(.4)
F0inB2	0	(2)	0	(0)	0	(0)	0	(0)	0	(1)	0	(2)
F0deB2	0	(1)	0	(0)	0	(5)	0	(0)	0	(3)	0	(3)
F0deA2	0	(.7)	0	(0)	0	(0)	0	(0)	0	(3)	0	(4)
F0inA2	0	(.3)	0	(0)	0	(0)	0	(0)	0	(0)	0	(2)
AF0inB2	0	(.4)	0	(0)	0	(0)	0	(0)	0	(.4)	0	(.6)
AF0deA2	0	(.3)	0	(0)	0	(0)	0	(0)	0	(.5)	0	(.4)
AF0deB2	0	(.3)	0	(0)	0	(.4)	0	(0)	0	(.2)	0	(.3)
AF0inA2	0	(.3)	0	(0)	0	(0)	0	(0)	0	(0)	0	(.4)
F0sno5B	0	(2)	0	(0)	0	(0)	0	(0)	0	(2)	1	(2)
F0sno10B	0	(4)	0	(0)	0	(0)	0	(0)	0	(4)	1	(4)
F0sno16B	0	(6)	0	(0)	0	(.7)	0	(0)	0	(5)	5	(6)
F0sno23B	0	(9)	0	(0)	0	(3)	0	(0)	0	(7)	6	(9)
F0sno31B	0	(12)	0	(0)	0	(6)	0	(0)	0	(10)	10	(11)
F0sno40B	0	(14)	0	(0)	0	(10)	0	(0)	0	(12)	17	(14)
F0sno5A	0	(2)	0	(0)	0	(1)	0	(0)	0	(2)	0	(2)
F0sno10A	0	(3)	0	(0)	0	(3)	0	(0)	0	(4)	4	(4)
F0sno16A	0	(5)	0	(0)	0	(5)	0	(0)	0	(6)	7	(6)
F0sno23A	0	(6)	0	(0)	0	(7)	0	(0)	0	(8)	12	(9)
F0sno31A	0	(7)	0	(0)	0	(10)	0	(0)	0	(10)	16	(12)
F0sno40A	0	(8)	0	(0)	0	(12)	0	(0)	0	(13)	20	(15)
F0smB	0	(9)	0	(0)	0	(9)	0	(0)	0	(6)	1	(13)
F0smA	0	(2)	0	(0)	0	(12)	0	(0)	0	(9)	0	(13)
F0intB	0	(8)	0	(0)	0	(13)	0	(0)	0	(6)	1	(14)
F0intA	0	(1)	0	(0)	0	(0)	0	(0)	0	(8)	0	(9)

Table 5.8: Median values (Md) and standard deviations (SD, in brackets) for the acoustic features in the **F0** domain for the boundary tones in the GToBI training material.

RMS												
Tone	**L-L%**		**L-H%**		**H-H%**		**H-L%**		**L-**		**H-**	
nr items	27		5		8		2		10		25	
	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD	Md	SD
RMinB	0	(.9)	.5	(.5)	0	(.7)	0	(0)	0	(.8)	1	(1)
RMdeB	3	(5)	2	(2)	4	(6)	4	(2)	1	(6)	0	(4)
RMinA	0	(.6)	0	(0)	0	(1)	.5	(.5)	.5	(2)	1	(3)
RMdeA	1	(1)	1	(0)	1	(1)	0	(0)	.5	(2)	0	(1)
ARMinB	0	(.6)	0.5	(.5)	0	(.6)	0	(0)	0	(3)	1	(1)
ARMdeA	1.09	(2)	1.17	(.1)	1.05	(.9)	0	(0)	0.57	(1)	0	(1)
ARMdeB	0.13	(.4)	0.06	(0)	0.05	(.2)	0.26	(.1)	0.1	(.3)	0	(.3)
ARMinA	0	(.5)	0	(0)	0	(.4)	0.49	(.4)	0.03	(.2)	0.09	(.4)
RMsm5B	0	(1)	.5	(.5)	0	(2)	0	(0)	0	(2)	1	(2)
RMsm10B	0	(1)	.5	(.5)	0	(3)	2	(2)	1	(3)	1	(2)
RMsm16B	0	(2)	3	(3)	.5	(5)	3	(3)	2	(5)	1	(3)
RMsm23B	0	(2)	3	(3)	.5	(6)	3	(3)	5	(7)	1	(5)
RMsm31B	1	(3)	3	(3)	.5	(6)	3	(3)	6	(9)	1	(6)
RMsm40B	1	(3)	3	(3)	.5	(7)	3	(2)	7	(12)	2	(8)
RMsm5A	4	(1)	5	(0)	4	(2)	3	(2)	2	(2)	1	(2)
RMsm10A	9	(3)	10	(0)	8	(4)	6	(5)	3	(4)	1	(2)
RMsm16A	15	(4)	15	(2)	14	(6)	9	(8)	7	(6)	2	(3)
RMsm23A	22	(6)	18	(5)	19	(8)	12	(11)	7	(8)	3	(5)
RMsm31A	30	(8)	22	(9)	25	(11)	17	(14)	9	(10)	4	(7)
RMsm40A	39	(11)	27	(14)	34	(14)	23	(18)	9	(12)	6	(9)
RMscB	3	(4)	1	(6)	2	(4)	8	(4)	3	(9)	2	(6)
RMscA	39	(14)	39	(16)	39	(16)	25	(15)	2	(12)	2	(12)
RMpsB	0	(2)	0	(1)	0	(13)	.5	(.5)	.5	(6)	0	(4)
RMpsA	9	(13)	9	(7)	10	(12)	5	(5)	1	(8)	5	(12)
RMsmrB	4	(9)	1	(6)	7	(8)	16	(14)	3	(11)	4	(6)
RMsmrA	30	(10)	29	(12)	30	(11)	19	(10)	4	(10)	5	(10)

Table 5.9: Median values (Md) and standard deviations (SD, in brackets) for the acoustic features in the RMS domain for the boundary tones in the GToBI training material.

accents and boundary tones were used to define the final definition of selection criteria. Therefore, the established values for the individual acoustic parameters are not directly transferred into the programs selection criteria, but expanded by additional knowledge from visual inspections.

5.4 Phonological Mapping

Since there is no one-to-one relation between these features and the symbolic, phonological labels, a mapping process had to be established between these two representations. In figure 5.1 (page 100) two arrows are depicted between the boxes including "Features of F0 and RMS" and "Phonological Mapping" to indicate that it is not solely a bottom-up process but also a structuring top-down process which is involved in this mapping procedure. The exact definition of this process is still subject of investigation and the present work intends to shed some light on it.

The mapping procedure evaluates feature combinations with respect to their agreement with predefined feature bundles for specific pitch accents. The mapping algorithm works as follows: each pitch accent is defined with a feature vector that includes filter values for each of the acoustic features. In dependence of the importance of the individual feature for the specific pitch accent, points are distributed and summed up. Therefore, each single frame will have an exact number of points which it received during the analysis of its acoustic features for each of the individual pitch accents and boundary tones. Pitch accent and boundary tone candidates are then subsequently selected by applying threshold values for each of the tones. Whenever an individual frame exceeds the threshold, it will be selected as candidate for the subsequent processing. The threshold values have been estimated by several tests, though they have not yet been optimized by numerical means.

The strategy chosen within the mapping algorithm uses a scoring system that gives positive points for parameter configurations that support the existence of a specific pitch accent or boundary tone and gives negative points for parameter configurations that do not. The use of a scoring system can be motivated with the better handling of the acoustic variability as compared to the 'hard decisions' within a computer program. Instead of defining threshold values as filter criteria and directly using them to select pitch accent candidates, the scoring system leaves more room for the sometimes drastically varying acoustic data and subsequently does not rule out possible candidates in an early decision step.

For each frame the scoring system chooses the tone with the highest score. Only those are subject of subsequent processing. After the pitch accent and boundary tone candidates have been selected, the phonological mapping procedure further checks their sequence and distance and writes the final output a label file with the type of prosodic event and its position in time.

The next section will deal with the question why there was no usage of HMM methods for this purposes and will argue for the advantages of the rule-based approach employed here.

5.5 Rule-Based vs. HMM

One of the questions that might have been appeared during the last sections is: why not choose a more automatic way of parameter assessment as in a HMM (Hidden Markov Model) based approach. HMMs can be used to model any time series and in the task presented here it would take a manually labeled corpus as input and would use statistical procedures to estimate the parameters necessary for recognition without knowing anything about the underlying structure of intonation.

The latter aspect makes the HMM approach unattractive for a system that tries not only to reach a good recognition accuracy, but also provide some insights in the underlying prosodic structure and its interaction with the acoustic features.[3] HMMs may result in "black-box" feelings, that is where some input is given to the black-box that processes it in some, nontransparent way and delivers an output. Processes within the black-box might be mathematically explainable, but do not give any insights about speech specific structures.

Furthermore, the rule-based approach presented in this work enables one to systematically change selection criteria and to check its consequences immediately. Changes at any level can be done easily without the necessity of any time consuming retraining. Each processing level, beginning from the calculation of acoustic features up to the phonological mapping procedure is transparent and may be analyzed and changed. During the development of the algorithm, it was very insightful to have the possibility of making changes at any level and observing its consequences on separate levels and of course on the final output.

Now that the architecture and the implementation of the ProsAlign model has been presented along the method and the results of the parameter assessment program the following chapter will deal with the evaluation of the program.

[3] See also the critique of Hidden Markov models in Reetz (1998, Chapter 2.1).

Chapter 6

Evaluation of the Program

6.1 Introduction

This chapter presents the evaluation of the data generated by the ProsAlign program. The evaluation is provided in two separate methods: (1) ProsAlign's output is compared to manually annotated labels in corpora used to train human labelers for the GToBI and the ToBI labeling instructions, and (2) ProsAlign is compared to four other human labelers who labeled one and the same corpus, thus including the variability existing between human labelers. The introduction gives an overview of the problem areas, especially the disagreements over the identity and position of tone labels annotated by different human labelers. Then the method of the first evaluation is explained in detail including the decision about allowable positional deviations between manually and automatically established labels. Afterwards the results are presented in overviews and in more detailed tables for both pitch accents and boundary tones. Finally, the discussion summarizes the results and reflects on them critically. Since the first evaluation included only comparisons with one manually labeled version of a corpus, it ignored the variability existing between different human labelers. In order to show ProsAlign's ability to act similarly to a human labeler would, a further evaluation was conducted. This evaluation compared the labels of 4 human labelers labeling an extended version of the GToBI training corpus with the ones produced by ProsAlign. Statistical methods already established for earlier inter-transcriber reliability studies (e.g. Silverman et al. 1992; Pitrelli et al. 1994; Grice et al. 1996; Syrdal & McGory 2000) were used in order to provide comparability to those.

At first glance, the method of evaluation seems obvious by using already manually annotated corpora as a reference and by comparing the automatically generated label files with the corresponding manually established ones. However, there are a number of problematic aspects when doing so. The disputed point here is that

- human labelers disagree about the position and type of labels.

Although there are studies that have shown that there is usually good agreement between trained labelers (e.g. Grice et al. 1996; Reyelt et al. 1996) an experienced labeler would never expect the prosodic annotations of two different human labelers to be absolutely identical. Often there are confusions between the following pairs of pitch accents

- H* and L+H*,
- L*+H and L+H*,
- L* and L*+H, and
- H* and H+!H* (Grice et al., 1996),

and regarding the boundary tones between

- L-L% and L-H% (Syrdal & McGory, 2000).

Since there exist different possible labelings for one and the same speech material, it is necessary for the evaluation to set the limits between allowable and impossible tone assignments. Opposed to the variability in position and type of prosodic labels introduced by humans, ProsAlign has the advantage to apply objective criteria and does not rely on human language intuition. Therefore, one of the fundamental advantages of an automatic procedure is the consistency in labeling prosodic events. When the program produces reliable results, it could unify the annotations of speech corpora.

In the first evaluation, ProsAlign has to show its performance compared to manually established labels. One way to reduce between-transcriber variations is to use training material from labeling instruction material that has been annotated by several labelers and where the final state is assumed to represent (1) a generally agreed way to label, and (2) prototypical cases of individual tone categories.

As a result of the above caveats, it was decided to take the training material from manually labeled corpora as reference for the evaluation. In order to get a reasonable basis for the evaluation and to enable the comparison of ProsAlign's performance on two different languages, two corpora of two different languages were chosen - GToBI and ToBI. GToBI[1] is a collection of sound files with accompanying prosodic label files for German (Grice & Benzmüller, 1997; Grice et al., to appear) It includes 40 prosodically labeled example utterances each provided with a corresponding soundfile and four label files for positions of pitch accents and boundary tones, word-boundaries, break-positions, and miscellaneous annotations according to the ToBI labeling conventions, but with adaptations specific for German. The

[1] See the presentation of this model in section 3.2.3. The GToBI training corpus is available under http://www.coli.uni-sb.de/phonetik/projects/Tobi/index_training.html.

utterances include altogether about 290 words and have different durations ranging from one word up to several words, some of them with intonation phrase breaks in between. The examples include a number of different speakers, both male and female, different recording qualities (different recording peak levels, background noises, cross talk) as well as different sampling rates (9600 and 16000 Hz). The text style ranges from parts of dialogs about routes from A to B (map task), parts of talks about appointments up to radio news excerpts. Speaking styles range from sloppy conversational to formal.

The American English ToBI corpus[2] includes about 150 examples (about 1680 words) of American English utterances. All of them are labeled according to the ToBI conventions (Beckman & Ayers, 1997). The recordings include several different male and female speakers as well as different recording qualities (background noises, low recording levels, sampling rates from 8000 Hz up to 20000 Hz) and also different speaking styles (fast, slow, emotional, hesitated, etc.). The utterances range from one-word sentences up to several words with several intermediate intonation phrase breaks. A wide range of text styles is captured starting from simple statements, repetitions with different intonation contours, information requests, parts of radio broadcasts and radio news, dialogs with and without hesitations up to flight information requests. Therefore, most of the examples do not include laboratory speech from controlled recordings in unechoic rooms and some of the examples are even difficult to understand.

Despite choosing corpora consisting of training material for the evaluation, two general problem areas remain, that is how to evaluate

1. differences in **label position in time**, and
2. differences in **tone identity**.

The first problem appears as follows:

- Manually established labels are often set at positions according to labeling conventions whereas the program relies on the existence of acoustic cues at a specific region in time. Since the algorithm does not know the position of phoneme-, syllable or word-boundaries it has to rely solely on these acoustic features. Therefore the positions of manually and automatically set labels might differ some period of time that is approximately of the size of a syllable. Though this period of time is not exactly determinable and it has to be decided what are the permissible variations.

The second problem comes up as follows:

[2]Downloadable from ftp://www.ling.ohio-state.edu/pub/TOBI/ame_wav_files/.

- The tone set of languages like German or English include tone contrasts like H* vs. L+H* and L* vs. L*+H. Although there are definitions of each single tone that should prevent mismatches labeling experience shows that there are often problematic cases in new material and that final decisions are often based on an inspection of larger passages of the described material. Here it has to be decided what counts as mismatch in tone type. Of course an obvious mismatch would be a H* in the manual label file whereas at the same position in the automatically established label file is a L*, but how to handle the other cases including cases with downstepped tones?

The next section addresses these two problem areas and lays out the method of evaluation.

6.2 First Evaluation

6.2.1 Method

There are 3 types of possible errors in the automatically generated label file:

1. **Missing value**, that is, there is no correspondent tone - neither in position nor in type - in the automatically generated label file as compared to the manually generated label file.

2. **Insertion**, that is, a tone in the automatically generated label file that has no correspondence - in position - in the manually established label file.

3. **Mismatch** in tone type, that is when there is a tone in the automatically generated label file at the same position as in the manually established label file but without correspondence in tone type, for instance a L* in the automatically generated label file whereas at the same position is a H* in the manually labeled file.

The other possibilities are:

1. **Perfect match** in time and tone type.

2. **Partial match** in tone type, that is a match in time and partial match in tone type, for instance a H* in the automatically generated label file whereas at the same position is a L+H* in the manually labeled file or vice versa. See table 6.1 for an overview of possible partial matches for the individual tones.

Tone	May be partial match for
H*	L+H*, *?, X*?, H+!H*, !H*
L+H*	H*, *?, X*?, H+!H*, !H*
H+!H*	H*, L+H*, *?, X*?, !H*
L*	L*+H, *?, X*?, L*+!H
L*+H	L*, *?, X*?
H+L*	L*, L*+H, *?, X*?
L-L%	H-L%, %, %?, X%?
H-L%	L-L%, %, %?, X%?
L-H%	H-H%, %, %?, X%?
H-H%	L-H%, %, %?, X%?

Table 6.1: List of possible partial matches for the individual boundary tones and pitch accents.

With respect to the last two possibilities one has to decide what amount of temporal offset between automatic and manual label is reasonable. The size of a syllable seems to be appropriate since the labeling conventions state that a pitch accent should be placed within the nucleus of the accented syllable. However, syllables are of very variable duration and may range from a few milliseconds up to several hundred milliseconds depending on the speaking rate, segmental material within the syllable, stress, etc.

Visual inspection and comparison of automatically and manually set labels reveals that the last mentioned aspect is crucial since often the automatically set labels are perfect matches in type but are often shifted towards the right or left of the corresponding manually set label and would therefore be counted as insertions or as missing values.

For the automatic analysis procedure it was decided to allow shifts in time position up to 61 ms left or right of the manually set labels which seemed to provide a reasonable amount of tolerance and at the same time avoiding the risk of mismatches of tones. The chosen interval is fairly restrictive[3] in order to get reliable results about the algorithms performance. Choosing larger intervals allows more distant tone positions and subsequently improves the outcome of the results.

To perform the evaluation, the F0 and RMS contours where calculated from all the speech files in the two corpora, using the *get_f0* program in version 1.14 from the ESPS/waves-tools (see page 99 for comments and references on this program). The default values of the program were used. All those F0 files were processed

[3] Taylor (1994, p. 100 f) chooses a different evaluation method by using penalties of 0.1 for every 10 ms of misaligned boundaries. He allows up to 300 ms which results in a penalty of 3.0 and equals the penalty given for an identity error in pitch accent labeling.

Tone	Subsumed Tones	
H*	H*	!H*
L+H*	L+H*	L+!H*
H+!H*	H+!H*	!H+!H*
L*	L*	
L*+H	L*+H	L*+H?
H+L*	H+L*	
L-L%	L-L%	
H-L%	H-L%	!H-L%
L-H%	L-H%	
H-H%	H-H%	!H-H%

Table 6.2: Subsumption of downstepped variants of tones into one class.

with the *ProsAlign* program in order to get the automatically calculated positions of pitch accents and boundary tones. The resulting label files were used to make the comparison with the manually established tones label files. In order to make this comparison automatically an evaluation program was developed. This program takes the manually produced label file as 'master' and checks whether there are instances of the five cases mentioned above in the corresponding automatically generated label file and writes the output in a file.

Since there was no separation of downstepped variants of tones these cases where combined into one class as depicted in table 6.2. Other cases as the six pitch accents and the four boundary tones are handled as described in table 6.3.

6.2.2 Results

6.2.2.1 Results for GToBI corpus

To get an overview of the procedure first a segment of the results of the GToBI corpus is presented in table 6.4. The table lists for some files in the GToBI corpus the results of the comparison of manual and automatic label file. For instance, the first file in table 6.4 presents the results for the file named 'august'. Here the manual labeling has in total 3 tones whereas the automatic one has 5. From these 5 tones are 2 perfect matches, no partial matches, 3 are insertions, and there are neither mismatches nor missing tones. Altogether 40 files are processed including a total sum of 175 tone labels in the manual label files whereas the automatic label files include 215 labels (see beginning of the second last row in table 6.4). This shows the tendency of the algorithm to annotate more labels than the human labelers are doing. When defining the number of automatically established labels as 100%

Entry	Handling	Description
*	partial match pitch accent	accented syllable
*?	partial match pitch accent	not sure whether this syllable is accented
X*?	partial match pitch accent	syllable accented but uncertain what accent type
%	partial match bound. tone	boundary tone but uncertain about type
%?	partial match bound. tone	boundary tone but uncertain what type
X%?	partial match bound. tone	boundary tone but uncertain what type
L-, H-	not counted	low and high phrase accent
%H	not counted	initial high pitch accent
X-?	not counted	phrase accent but uncertain what type
-?	not counted	not sure whether there is a phrase accent
-	not counted	phrase accent but no type assigned
HiF0	not counted	F0 max. assoc. but not strictly aligned with high accnt.
>	not counted	early F0 event associated with a L or H pitch accent
<	not counted	late F0 event associated with a L or H pitch accent
%r	not counted	left edge of inton. phrase that begins after hesitation

Table 6.3: Handling of entries in the label files other than the six pitch accents and the five boundary tones.

reference level, the percentage of perfect matches is 43% and the number of partial matches is 13%. This means that 43 + 13 = 56% of the manually established labels are detected by the algorithm, although not all of them perfectly. The number of insertions is 39% from the 215 automatically set labels indicating that the algorithm labels much more intonational events than the human labelers are doing. However, this number as well as the number of missing tones (total number: 44 of 175 = 25% of all manually labeled tones) will be shown in their true light in the second evaluation presented in section 6.3.

Looking at the overall results one can state that the number of mismatches is fairly small. Only about 12 labels out of 215 (i.e. 6%) are mismatches. A closer inspection reveals that the main source of these mismatches results from boundary tone mismatches (8 cases out of a total of 47 boundary tones) and not from pitch accent mismatches (4 cases out of a total of 168 pitch accents). This implies that the intonation phrase final F0 movements are more often misleading, probably resulting from final laryngealizations, and therefore shows the limitations of the mechanism in detecting those cases. Though there is still room for improvement regarding the detection of boundary tones within ProsAlign, another possible solution for this problem could be an end-of-phrase-detector specifically designed for this purpose. The inclusion of further acoustic cues like spectral features might be considered for such a detector.

GToBI	**Nr of tones in**		**Nr of tones from auto that are**				
File	**hand**	**auto**	**perfect**	**partial**	**insertn**	**mismtch**	**missing**
august	3	5	2	0	3	0	0
blaue	2	2	2	0	0	0	0
blumen2	3	3	2	1	0	0	0
dina4	6	6	2	2	2	0	2
laengs	4	4	1	3	0	0	0
pension	5	6	3	0	3	0	2
...	...	...	...	...	...	...	...
Sum: 40	175	215	92	27	84	12	44
% of auto		100%	43%	13%	39%	6%	25% of hand

Table 6.4: Segment of the results of the evaluation of program **ProsAlign** for the GToBI corpus. The leftmost column shows the name of the individual file. Altogether there where 40 files processed including a total sum of 175 tone labels in the manually labeled set (see column entitled "hand") whereas the total sum of automatically annotated labels was 215 (next column entitled "auto"). The next 4 columns present the numbers of perfect matches (perfect), partial matches (partial), insertions (insertn), and the mismatches (mismtch). The last column shows the missing tones (missing). Finally the last row shows the relative percentage of each column when taking the total sum of automatically detected tones as 100% reference (except for the missing tones where the total number of manually detected tones is the 100% reference).

Of the four cases of pitch accent mismatches, three were a mismatch between a automatically labeled L*+H vs. a manually labeled H*. The F0 contours in all the three cases are rising during the stressed syllable and ProsAlign probably decides to annotate a L*+H accent, because the maximum in the simultaneous RMS amplitude appears more than 50 ms before the maximum of the rise. Adjusting the feature weights slightly differently could probably eliminate these cases. However, it remains unclear whether this readjustment of feature weights will introduce problems in the other direction and will create mismatches between automatically labeled H*'s and manually annotated L*+H's. Future implementations of ProsAlign are intended to include objectively optimized feature weights. The remaining case of mismatch includes a manual L* placed only 10 ms after a local maximum in F0 and a automatically labeled L+H* 20 ms before, whereas the L* was labeled about 100 ms before in the beginning of the rise. In the latter case, the algorithm was clearly guided by the local maximum in F0. The latter could be solved by using a deletion rule for (L+)H* labels which occur within a certain time interval after a L* accent that has high confidence ratings.

This first overview of the results shows that there is not a significant number of mismatches between human and automatic labels and it also shows that the algorithm covers at least about 56% of the manually established labels. Importantly, the automatically established labels are most often annotated at positions were at least the

GToBI	**Nr of tones in**		**Nr of tones from auto that are**				
	hand	**auto**	**perfect**	**partial**	**insertn**	**mismtch**	**missing**
Original:	175	215	92	27	84	12	44
% of auto		100%	43%	13%	39%	6%	25% of hand
All H*:	175	215	35	43	84	53	44
% of auto		100%	16%	20%	39%	25%	25% of hand
Random:	175	215	5	10	163	36	124
% of auto		100%	2%	5%	76%	17%	71% of hand

Table 6.5: Comparison of the evaluation results for the GToBI corpus for different scenarios. The row starting with "Original" shows the results from the original evaluation. The one starting with "All H*" shows the results when all the automatically labeled tones are changed to "H*" (which is the label category with the highest occurrence in the GToBI corpus) while keeping the number and positions of the tones the same. And the row beginning with "Random" shows the results when both label type as well as label position is marked in a random way while keeping the overall number of tones the same as in the original evaluation. Cf. table *6.4*.

author of this thesis could think of providing a tone label and rarely at positions that are not linguistically motivated. To estimate the performance of ProsAlign, it was compared to two other scenarios: (1) when changing all the label types to "H*", which is the category with the highest occurrence in GToBI, and (2) when changing the type and position of the labels in a random way. The results are depicted in table 6.5. In the first scenario, the number of perfect matches decreases significantly (from 43% to 16%), the number of partial matches of course increases (from originally 13% to 20%) and the number of mismatches increases to 25% from originally only 6%. In the second scenario it becomes even more obvious that ProsAlign performs quite well. Here, the number of perfect matches drops to 2%, the number of partial matches falls to 5% and the insertions increase drastically to 76% and also the number of mismatches to 17%; and the number of missing tones also increases (from 25% to 71%) as a result of the different positions of labels.

Results for pitch accents

In table 6.6, a confusion table for the manually and automatically labeled pitch accents in the GToBI corpus is illustrated. The table shows the number of perfect and partial matches as well as the number of insertions, mismatches and missing tones for each individual pitch accent. Altogether there are 124 pitch accent labels in the manually labeled files and 168 labels in the automatically produced ones. This indicates the algorithm's tendency to insert more labels than the human la-

Pitch Accents in GToBI-corpus							
	Manually labeled (incl. downstepped variants)						
	H*	**L+H***	**H+!H***	**L***	**H+L***	**L*+H**	
Total sum	63	28	8	11	6	8	
Detected (Nr)							Insertns
H* (79)	49%, 31	25%, 7	63%, 5	-	-	-	46%, 36
L+H* (47)	13%, 8	64%, 18	-	-	-	13%, 1	43%, 20
H+!H* (12)	2%, 1	-	25%, 2	-	-	13%, 1	67%, 8
L* (14)	-	-	-	72%, 8	-	13%, 1	35%, 5
H+L* (2)	-	-	-	-	17%, 1	-	50%, 1
L*+H (14)	3%, 2	-	-	9%, 1	-	25%, 2	64%, 9
Not detected	33%, 21	11%, 3	13%, 1	18%, 2	83%, 5	38%, 3	

Table 6.6: Confusion table of manually labeled vs. detected pitch accents for the GToBI corpus. First the percentage and then comma-separated the absolute number of items in each individual class are depicted. Altogether there were 40 files processed including a total sum of (63 + 28 + 8 + 11 + 6 + 8 =) 124 pitch accent labels in the manually labeled set whereas the total sum of automatically produced labels was (79 + 47 + 12 + 14 + 2 + 14 =) 168.

belers which is also reflected in the number of insertions (total sum: 79 of 168 = 47% of the automatically annotated labels). The confusion table also shows that the recognition accuracy is different for the individual pitch accents, ranging from 72% for the L* tones to only 17% for the H+L* tones. Not surprisingly, 25% of the accents manually labeled as L+H* are labeled as H* by ProsAlign. Even 63% of the manually labeled H+!H*'s are labeled as H* by the algorithm and only 25% of the H+!H*'s are labeled perfectly. 13% of the accents originally labeled as H* are marked as L+H* here, once again indicating the difficulty of separating the H*'s from the L+H*'s. However, the confusions are mostly within the same class of pitch accents, that is either in the high or in the low class and confusions of this type are also common for human labelers (cf. page 136).

With 72% the L*'s have the highest recognition accuracy, whereas only 1 out of 6 H+L*'s is detected (though the small number of this accents prohibits a serious statistical treatment, but they were included for reasons of completeness) and neither are there partial matches by the other low tone categories indicating that this accent includes very diverse acoustic features. The L*+H category is recognized by only 25% perfect, one partial match for a L*.

Altogether the algorithm separates the two categories high vs. low well and confusions are mostly in between the same (high or low) class. The number of tones that are not detected is about 30% for H*'s, and about 12% for L+H*'s and H+!H*'s. Therefore, about 20% of the tones manually labeled as high are not detected by the algorithm. The number of insertions is larger, namely about 45% insertions in both

the H* and L+H*-categories. 67% insertions in the H+!H* category indicate some problems with this class. Here it seems that the algorithm finds many more positions that fit the description of H+!H* tones, but are not labeled by human labelers. At this point it could be helpful to formulate additional phonological deletion rules based on the existence of neighboring accents and boundary tones. The L* has only 35% insertions whereas it is 64% in the L*+H category. The low performance in the H+L* category indicates probably insufficient features or very diverse feature values, but could also indicate problems with the phonological status of this class. Although the small number of cases in this class prohibits serious statements about the phonological status of the H+L* category, it is nevertheless possible to use ProsAlign as a verification tool for the postulated categories within a phonological theory of intonational structure in a given language. Whenever there are consistently poor recognition results for a specific class, this indicates that the class in itself was stipulated incorrectly within the phonological model.

As for insertion errors, there is one type which occurs fairly often when labeling a rise in F0. The rise is labeled by ProsAlign early in the rising part either as L* or L*+H (which comforms often to the manual labeling) but later in the rise towards the peak of it another high pitch accent label is inserted (see e.g. figures 2 and 6 in Appendix A). These cases might be avoided by applying an additional deletion rule for the high pitch accents based on the score of the individual accents or as phonological deletion rule.

Results for boundary tones

Table 6.7 shows a confusion table for the manually and automatically labeled boundary tones in the GToBI corpus. L-L% boundary tones are fairly well detected (74%), whereas the other low boundary tone preceded by a high phrase accent is recognized in 2 of 4 cases that is 50% (though the small number of cases does not allow serious statements here). The high boundary tones are also not very well recognized: 33% of the L-H%'s and 50% of the H-H%'s. The number of mismatches especially between L-L% and H-H% are: 3 out of 31 manually labeled L-L% are labeled as H-H% by the algorithm and the other way around 3 out of 10 manually labeled H-H% are labelled as L-L% by the program. It seems that the procedure to detect final rises and falls does not reliably differentiate between those two cases. It is highly probable that the mechanism to detect erroneous F0 values is less effective for the boundary tones than it is for the pitch accents, since the latter ones include many more 'context' material with respect to F0, where it is easier to base decisions on, whereas the F0 track usually (however not necessarily) ends abruptly at the end of an intonation phrase. Other reasons for undetected boundary tones are an abrupt cutting off of the signal at the end of the intonation phrase where there is no contextual material to base a decision on, and cases where no clear pause appears after the intonation phrase and where mainly segmental features like pre-final lengthening are relevant.

Boundary Tones in GToBI-corpus					
	Manually labeled (incl. downstepped variants)				
	L-L%	**H-L%**	**L-H%**	**H-H%**	
Total sum	31	4	6	10	
Detected (Nr)					Insertns
L-L% (32)	74%, 23	50%, 2	17%, 1	30%, 3	9%, 3
H-L% (0)	-	-	-	-	-
L-H% (5)	3%, 1	-	33%, 2	10%, 1	20%, 1
H-H% (10)	10%, 3	-	17%, 1	50%, 5	10%, 1
Not detected	13%, 4	50%, 2	33%, 2	10%, 1	

Table 6.7: Confusion table of manually labeled vs. detected boundary tones for the GToBI corpus. First the percentage and then the absolute number of items in each individual class are given separated by a comma. Altogether, 40 files were processed including a total sum of (31 + 4 + 6 + 10 =) 51 boundary tone labels in the manually labeled set, whereas the total sum of automatically produced labels was (32 + 0 + 5 + 10 =) 47.

Summary

What are the characteristics of the cases that show perfect matches? (a) the presence of obvious acoustic cues for the individual tone category, (b) no disturbing influences of laryngealization or background noise, (c) correspondence with the feature evaluation criteria defined in the scoring system, (d) no disturbing contextual influences that could deselect candidates in the phonological mapping process.

What are the characteristics of the partial matches? (a) often cases manually labeled as L+H* are labeled by ProsAlign as H* (7 of totally 28 manually labeled L+H* accents). 8 out of altogether 47 manually as H* labeled pitch accents were recognized by ProsAlign as L+H*. The latter are mismatches that occur also often in manual labeling, as already noted in the beginning of this chapter. 5 out of a total of 8 manually as H+!H* transcribed accents are labeled automatically as H*. Regarding the boundary tones, there are 2 out of 4 cases manually labeled H-L% accents that became automatically annotated as L-L%.

Most of the insertions are H* labels (36) followed by L+H* (20). Also a number of H+!H* (8) and L*+H (9) labels were inserted by the algorithm. When checking the H* insertions within the corresponding speech file and the time aligned F0 and RMS tracks it often appeared to the author that these are 'possible' accent positions. In many cases, the perceptual impression confirmed the placement of a label and very seldom labels were placed in positions where there were absolutely no perceptual impressions of an accent.

Regarding the mismatches, there are only four in the automatically labeled pitch accents. Manually labeled H* was labeled automatically as L*+H and manually

L*+H was labeled automatically as L+H*. In both cases, ProsAlign labeled the rising F0 movement *both* L*+H at the beginning of the rise *and* L+H* towards the end of it. In this case, a deletion rule could at least eliminate one of the labels, though the decision for one or the other category might still not always coincide with the manually labeled version, since ProsAlign does not know the segment boundaries.

The percentage of mismatches is higher for the boundary tones: in 3 cases a manually labled L-L% tone was automatically labeled as H-H%. These cases had final F0 rising movements which were not realistic (faulty F0 calculations). Here, the separation of faulty and perceptually important F0 movements did not work. On the other hand, three boundary tones manually labeled as H-H% were labeled by ProsAlign as L-L%, which indicates that there are still difficulties in the reliable detection of final rising movements.

In summary, a major source of errors is not the algorithm to place and assign the labels, but are caused by faulty values of the F0 tracker. Human labelers can use their auditory impressions which are not subject to this error source.

After the performance of ProsAlign has been shown by using the GToBI corpus as basis, it was subsequently tested on a new corpus, to see how the program performs on material not included in the analysis of acoustic features and from a different language. Therefore, the next section describes the results of ProsAlign's performance for the American English ToBI training corpus. Here again, ProsAlign is compared to manually labeled reference data.

6.2.2.2 Results for ToBI corpus

Since it is claimed that the algorithm works for any language after adapting the set of possible phonological candidates, the ToBI training corpus (see page 137) including American English speech material was chosen as the new corpus on which to test the algorithm. Table 6.8 shows an excerpt of the results for the ToBI corpus. Altogether 158 files were processed including a total number of 1045 tone labels. Statistically astonishing, the algorithm here produces slightly less tone labels (1006), whereas it was the other way around in the GToBI corpus. The last result could be a consequence of the subtle differences in the acoustic properties of German and American English tones. The number of perfect matches (38%) is slightly smaller than in the GToBI corpus (43%), but the number of partial matches is clearly higher (25%), partly a result of the larger number (47) of unsure tones ([X]*?, cf. 6.9), but possibly again indicating that there are some differences in the subtle set of acoustic features of the individual pitch accents. The number of insertions is smaller in the ToBI corpus (30% as compared to 39% in GToBI). However, the number of missing tones increased (32% as compared to 25%) in the ToBI corpus, once again indicating that the set of acoustic features might have to be adapted to American English. However, when adding up the number of perfect

ToBI	**Nr of tones in**		**Nr of tones from auto that are**				
File	**hand**	**auto**	**perfect**	**partial**	**insertn**	**mismtch**	**missing**
howto	6	5	3	0	1	1	1
mother3	7	10	4	2	4	0	1
really1	9	8	3	2	1	2	2
wellies1	3	3	1	1	0	1	0
word	3	5	1	1	3	0	1
yellow3	4	4	1	2	1	0	1
...	...	...	...	...	...	...	...
Sum: 158	1045	1006	387	253	299	67	338
% of auto		100%	38%	25%	30%	7%	32% of hand

Table 6.8: Segment of the results of the evaluation of program **ProsAlign** for the ToBI corpus. The leftmost column shows the name of the individual file. Altogether there were 158 files processed including a total sum of 1045 tone labels in the manually labeled set (see column entitled "hand") whereas the total sum of automatically produced labels was 1006 (next column entitled "auto"). The next 4 columns present the numbers of perfect matches (perfect), partial matches (partial), insertions (insertn), and mismatches (mismtch). The last column shows the missing tones (missing). Finally the last row shows the relative percentage of each column when taking the total sum of automatically detected tones as 100% reference (except for the missing tones where the total sum of manually detected tones is the 100% reference).

matches and the number of partial matches (38% + 25%) the result is 63% which is even higher than the 56% overall recognition rate for the GToBI corpus. The latter shows ProsAlign's ability to be used for American English speech as well.

Table 6.9 shows a confusion table for the comparison of manually and automatically labeled pitch accents for the ToBI corpus. As was the case for the GToBI corpus there are also slightly more automatically transcribed labels (784) than manually ones (739). The detection rate of H* pitch accents is slightly smaller in the ToBI-corpus (44%) than in the GToBI corpus (49%). L+H*'s are also less reliably marked (ToBI: 39% vs. 64% in GToBI) and also often labeled as H*'s (44%). 67% of manually as H+!H*'s labeled accents are detected as H*'s by the algorithm similar to the results in the GToBI corpus.

The low pitch accents are much more poorly detected (contrary to the GToBI corpus where the L* category was detected best): only 15% in the L* cases and only 17% in the L*+H cases. Although the overall number of mismatches is very small there are 8 cases where a manually labeled L* was detected as H* by the algorithm and 3 cases where a manually labeled L*+H was detected also as H* by the algorithm. Whereas the latter cases are explicable by the distinct low-high F0 movement in L*+H cases where the algorithm decided to prefer the high pitch accent label, the mismatches between L* and H* are unexpected.

Pitch Accents in ToBI-Corpus							
	Manually labeled (incl. downstepped variants)						
	H*	**L+H***	**H+!H***	**L***	**L*+H**	**[X]*?**	
Total sum	415	156	12	97	12	47	
Detect (Nr)							Insertns
H* (436)	44%, 181	44%, 68	67%, 8	8%, 8	25%, 3	36%, 17	35%, 151
L+H* (246)	24%, 100	39%, 61	-	3%, 3	-	13%, 6	31%, 76
H+!H* (38)	3%, 13	5%, 7	17%, 2	1%, 1	-	4%, 2	34%, 13
L* (35)	0.5%, 2	-	-	15%, 15	17%, 2	-	46%, 16
L*+H (29)	0.7%, 3	-	-	12%, 12	-	-	48%, 14
Not detected	28%, 116	13%, 20	17%, 2	60%, 58	58%, 7	47%, 22	

Table 6.9: Confusion table of the results of the evaluation of program **ProsAlign** for the individual pitch accents in the ToBI corpus. First the percentage and then the absolute number of items in each individual class are given separated by a comma. Altogether there where 158 files processed including a total sum of (415 + 156 + 12 + 97 + 12 + 47 =) 739 tone labels in the manually labeled set whereas the total sum of automatically produced labels was (436 + 246 + 38 + 35 + 29 =) 784.

Most of the cases manually labeled as unsafe ([X]*?) were marked as one of the high pitch accents: 36% H*, 13% L+H*, and 4% H+!H*, indicating that these cases include acoustic features of high pitch accents.

The boundary tones in the ToBI corpus are generally less reliably detected than in the GToBI corpus. Also only for the boundary tones in the ToBI corpus are there more manually than automatically produced labels (which was the other way around in the GToBI corpus): 222 automatically annotated vs. 306 manually transcribed boundary tones whereas for the pitch accents the relation is: 784 automatically set vs. 739 manually annotated pitch accents. This indicates some problems with the boundary tone detection in the American English sentences. It could be partly explained by the larger number of noisy recordings but is probably also a result of the more longer utterances in the ToBI corpus. Moreover, the boundary tones detection algorithm is designed in a way that prevents labels from beeing set at positions in running speech. One of the crucial acoustic parameters is the duration of pause after an utterance along the constancy of low amplitude within this pause. Whenever there is noise or other effects within the pause, there is a high likelihood that the algorithm does not put a boundary label. However, there are also cases of final boundary tones that do not have pauses afterwards. These cases are usually marked with other acoustic features like phrase final lengthening or final lowering of the F0 curve. However, since there was no recognition of segmental content nor of the speakers pitch range this features could not be used. Furthermore, final lowering of the F0 curve is often not distinguishable from other not finality marking lowerings in the F0 track and is therefore not easily usable as

Boundary Tones in ToBI-corpus					
	Manually labeled (incl.downstepped variants)				
	L-L%	**H-L%**	**L-H%**	**H-H%**	
Total sum	196	19	46	45	
Detected (Nr)					Insertns
L-L% (151)	53%, 105	37%, 7	22%, 10	18%, 8	14%, 21
H-L% (0)	-	0%, 0	-	-	-
L-H% (26)	5%, 10	11%, 2	13%, 6	9%, 4	15%, 4
H-H% (45)	7%, 13	21%, 4	15%, 7	38%, 17	9%, 4
Not detected	35%, 68	32%, 6	50%, 23	36%, 16	

Table 6.10: Confusion table of the results of the evaluation of program **ProsAlign** for the individual boundary tones in the ToBI corpus. First the percentage and then the absolute number of items in each individual class are given separated by a comma. Altogether there where 153 files processed including a total sum of (196 + 19 + 46 + 45=) 306 boundary tone labels in the manually labeled set whereas the total sum of automatically produced labels was (151 + 0 + 26 + 45 =) 222.

reliable selection criteria. An example for a missing final (high) boundary tone is shown in example "jam1" in appendix A (on page 180), here probably a result of the cut-off of the speech file very close at the end of the last word and the labeling of the final rise as L+H* accent. In summary, the overall detection accuracy was even higher than for the GToBI corpus. About 63% of the manually labeled tones are recognized by ProsAlign. The diverse speech material is certainly a challenge for the program.

Although the labels in both corpora used for the evaluation are so called training material which should include prototypical cases of each individual tone, there is no coverage of the variability between human labelers. Therefore, the results of the first evaluation have to be judged with respect to the general problem of correspondency of different human labelers. Since it is an everyday experience in prosodic research that there are often large differences between human labelers when labeling one and the same speech material especially in tonal identity as stated by Syrdal & McGory in their study about "Inter-transcriber reliability of ToBI prosodic labeling":

> "Thus while transcribers agree very well on whether or not a word is prominent or whether or not a phrase boundary follows it, they often do not agree on the identity of the specific tone involved." (Syrdal & McGory, 2000, p. 238).

If the algorithm were to act just as another human labeler would, then this would be a proof of its viability. To conduct such a test, another evaluation procedure

was designed using the labels of four different human labelers who labeled the complete GToBI training corpus plus some additional speech material. The method and result of this study will be presented below.

6.3 Second Evaluation

In order to estimate the program's performance as compared to more than one human labeler it was compared with four human human labelers. The goal of this study was twofold: (1) evaluation of the reliability among human labelers, when using the GToBI labeling conventions, and (2) evaluation of the reliability of ProsAlign as compared to the reliability of the human labelers. An extended version of the GToBI training corpus was used as basis. Word level agreement in pitch accents, phrase accents and boundary tones were analyzed by using methods already established for earlier studies of inter labeler reliability of ToBI prosodic labeling (Silverman et al. 1992; Pitrelli et al. 1994; Grice et al. 1996; Syrdal & McGory 2000). These methods were used to provide comparability with them, but do not include the fine granularity and detailed treatment of each label category as the method used for the first evaluation. Inter labeler reliability rates for human labelers only and for human labelers plus automatic labeler are presented. When adding the automatically produced GToBI labels the overall inter labeler reliability rate remained nearly at the same level as for the human labelers only, therefore showing the quality of the automatically produced labels.

6.3.1 Method

6.3.1.1 Corpus

The corpus consisted of the full GToBI training corpus plus some additional material. The additional material contains mainly the recording of one female speaker reporting an episode during her stay in London. This additional material had not earlier been included in the acoustic feature analysis and was therefore new to ProsAlign. The 108 utterances include altogether about 720 words and are of different duration ranging from one word up to several words, some of them having intonation phrase breaks in between. Four labelers (L1, L2, L3, L4) labeled the material according to the GToBI guidelines except for break indices. The speech quality of the recordings is changing and there are a number of different speakers both male and female. The labelers started to learn GToBI labeling and therefore had no prior experience with this transcription method.

6.3.1.2 Statistical treatment

The statistical treatment of the label data followed previous studies about inter-transcriber reliability (Silverman et al. 1992; Pitrelli et al. 1994; Grice et al. 1996; Syrdal & McGory 2000) in order to provide comparability to those. These studies use the so-called "transcriber-pair-word" as basic unit for measuring agreement between labelers. It is a "comparison of the labels that two particular transcribers placed on one particular word or at one particular word boundary in the database. The measure of inter-transcriber consistency is then the percentage of transcriber-pair-words exhibiting agreement on a particular element in the transcription." (Pitrelli et al., 1994, p. 3). In the present study there were four labelers which results in six possible pairwise comparisons (L1-L2, L1-L3, L1-L4, L2-L3, L2-L4, L3-L4). Transcriber-pair-word agreement is 50% when three of four labelers agree on a label for a word. If two of four labelers agree on the transcription of a word, pairwise agreement is 17%.

6.3.2 Results

The 108 utterances resulted in (720 'words' x 6 =) 4320 transcriber-pair-words. In the rare cases when there were more than one pitch accent on a word, its syllables were used as basic analysis unit.

6.3.2.1 Results for human labelers only

Pitch accents

Overall agreement (the percentage of the total number of transcriber-pair-words in the corpus for which there was agreement) was 58%. This number counts only identical pitch accents as agreements, that is, !H* is unequal to H*, etc., which means that upstepped or downstepped variants are taken as separate entities. With relaxed criteria (upstepped and downstepped variants merged to their base category) the number of agreements increased to 61%. These numbers are smaller than the results published in earlier studies. For instance, for overall pitch accent agreement it was 71% in the GToBI study (Grice et al., 1996) as well as in the EToBI study (Syrdal & McGory, 2000) and 68% in the earlier EToBI study (Pitrelli et al., 1994). The reduced agreement rates are probably mainly a consequence of the relatively inexperienced labelers and secondarily of the diverse speech material.

Agreement for presence versus absence of pitch accents (pooled across accent categories) is 77%. Agreement on individual pitch accents is listed in table 6.11. This table calculates percent agreement of individual accents based on the ratio of agreeing transcriber-pair-words and the maximal number of transcriber-pair-words in cases which include at least one word with a label of the tone under inspection, that is cases without any tone label were not included in this table.

PA	Nr	% agree
0	1673	64.5
H*	341	30.8
!H*	143	30.0
^H*	68	25.8
H* relax	552	38.3
L+H*	89	24.4
H*+L	42	21.7
H+!H*	14	31.9
L*	313	33.3
L*+H	113	22.4
H+L*	33	22.4

Table 6.11: Percent pairwise labeler agreement of individual pitch accents (PA). The column entitled "Nr" shows the total number of each pitch accent category. "0" means unaccented. "H* relax" includes both !H* and ^H* merged with H*.

BT	Nr	% agree
H%	193	57.7
^H%	27	48.7
!H%	15	32.1
L%	192	69.9
%H	29	47.1

Table 6.12: Percent pairwise labeler agreement of individual boundary tones (BT). Column entitled "Nr" shows the total number of each boundary tone category.

Edge tones

Overall agreement on phrase accents was 89%. Overall agreement on boundary tones was 93%; for presence vs. absence of boundary tones 96%. When combining phrase accents and boundary tones into edge tones overall agreement is 89% on exact edge tone and 93% agreement on existence or non-existence of edge tones. Table 6.12 shows the agreement results for individual boundary tones once again based on the ratio of agreeing transcriber-pair-words and the maximal number of transcriber-pair-words in cases which include at least one word labeled with one of the single edge tones.

	exact PA	PA	exact BT	ET
L1	60.5	79.6	91.8	88.3
L2	54.6	74.2	93.1	89.2
L3	60.6	77.5	94.3	91.1
L4	57.3	83.9	91.7	86.0

Table 6.13: Consistency of individual labelers (L1-L4) represented in percent agreement of each labeler with the other three. "exact PA" means agreement on identical pitch accents whereas "PA" means agreement solely on the existence or non-existence of a pitch accent. "exact BT" means agreement on identical boundary tones and "ET" means agreement solely on the existence or non-existence of an edge tone (merging phrase accents and boundary tones).

Consistency of labelers

Consistency of individual labelers was measured by the percent of label agreement of each individual labeler with the other three ones. Table 6.13 shows the results for each individual labeler separated between pitch accent and boundary tone agreement.

6.3.2.2 Results with ProsAlign

In order to estimate the quality of the automatically produced GToBI labels the same tests as described in the section before were conducted but with a "circular exchange" of one of the human labelers transcriptions with those provided by ProsAlign. This results in the following four comparisons (LP = labeler ProsAlign): LP $\rightarrow$ L234; LP $\rightarrow$ L134; LP $\rightarrow$ L124; LP $\rightarrow$ L123, that is, in the first test the transcriptions of labeler 1 were replaced by those from ProsAlign, in the second test were the transcriptions from labeler 2 replaced by those from ProsAlign, etc. Once again inter-transcriber reliability was estimated by calculating the number of agreeing transcriber-pair-words.

By comparing the automatically produced labels in this way with the manually produced ones, it is possible to differentiate the variability between human labelers and the variability introduced by ProsAlign. Therefore, this method allows a better evaluation of the reliability of automatically produced labels than comparing them only with one manually established transcription as was the case in the first evaluation.

Interestingly, the overall agreement rates were nearly similar to the ones calculated for the human labelers only, which shows the quality of the automatically produced labels.

Test	exact PA	merged PA	PA
L1234	58.2	61.1	76.9
LP234	53.3	56.0	72.4
L1P34	52.9	56.7	72.8
L12P4	57.7	60.3	75.6
L123P	54.9	57.5	72.8

Table 6.14: Consistency of individual human labelers (L1234) with ProsAlign (LP) represented in percent agreement of each labeler with the other three. "Exact PA" means agreement on exact pitch accent, "merged PA" includes upstepped and downstepped variants merged with the main category, "PA" means percent agreement whether there is or is not a pitch accent.

Pitch accents

In table 6.14 the consistency of individual human labelers with ProsAlign is compared in percent agreement of each labeler with the other three. The row beginning with "L1234" includes the consistency in percent agreement for human labelers only. The table shows that the replacement of one of the human labelers affects the overall consistency ratings only slightly. Though the percent agreement rating drops for about 4% for the replacement of speakers 1, 2, and 4 it stays nearly at the same level when labeler 3 is replaced by ProsAlign. These comparisons show that ProsAlign acts almost similar as a further human labeler with regard to the consistency ratings between them and therefore shows the quality of the labels produced by ProsAlign.

Comparison of agreement values for individual pitch accents are depicted in table 6.15. The upstepped ^H* pitch accent is not included here since ProsAlign does not currently produce this label.

Results for edge tones

In table 6.16 the consistency of individual human labelers regarding boundary tones and edge tones are depicted. Edge tones include both phrase accents and boundary tones merged. The overall agreement ratings for exact boundary tones are nearly similar between the test with human labelers only and the other tests where one of the human labelers was replaced by ProsAlign. The same holds for the edge tones, but here there is a slight drop in the agreement ratings as a result of the small amount of monotonal phrase accent labels assigned by ProsAlign. These results show once again the reliability of the automatically produced labels. Table 6.17 shows the percent of agreement for the two boundary tones H% and L%.

PA	L1234	LP234	L1P34	L12P4	L123P
0	64.5	59.0	58.4	64.0	60.1
H*	30.8	27.7	30.1	31.3	29.4
!H*	30.0	24.4	27.6	24.1	25.8
L+H*	24.4	24.3	22.2	27.0	23.4
H*+L	21.7	14.2	20.0	14.8	16.7
H+!H*	31.9	21.2	22.2	26.7	26.7
L*	33.3	26.3	27.2	25.7	29.0
L*+H	22.4	21.0	19.6	20.7	23.4
H+L*	22.4	18.0	17.4	18.1	17.5

Table 6.15: Percent pairwise labeler agreement of individual pitch accents of human labelers only (L1234) opposed to the ones when one of the human labelers is replaced by ProsAlign (LP).

Test	exact BT	BT	exact ET	ET
L1234	92.7	96.1	88.6	93.5
LP234	90.7	94.4	87.0	92.8
L1P34	89.1	94.0	85.2	92.7
L12P4	90.0	94.8	86.6	93.7
L123P	91.0	95.7	89.0	95.0

Table 6.16: Consistency of individual human labelers (L1234) with ProsAlign (LP) represented in percent agreement of each labeler with the other three. "Exact BT" means agreement on exact boundary tone, "BT" means percent agreement whether there is or is not a boundary tone. The same classification are used for edge tones (ET), which include phrase accents and boundary tones.

BT	L1234	LP234	L1P34	L12P4	L123P
H%	57.7	45.1	32.4	33.3	36.7
L%	69.9	56.5	50.0	50.3	55.8

Table 6.17: Percent pairwise labeler agreement of H% and L% boundary tones of human labelers only (L1234) opposed to the ones produced when one of the human labelers is replaced by ProsAlign (LP).

ProsAlign currently produces no downstepped or upstepped variants of boundary tones.

The analysis of the boundary tones showed that the automatically produced labels are fully compatible to manually established ones.

Consistency of labelers with ProsAlign

Table 6.18 shows the comparisons of consistency ratings between each individual human labeler with the other three and between ProsAlign and the other three human labelers. Results are separated for pitch accents and boundary tones. Agreement for exact pitch accent is slightly reduced by about 1% up to 9% when ProsAlign replaces one of the human labelers. Agreement for existence or non-existence of pitch accents is also only slightly reduced when ProsAlign is introduced, though the differences are larger for comparisons L4-L123 vs. LP-L123 and L1-L234 vs. LP-L234, but in general there is no significant drop in consistency ratings.

Regarding the boundary tones the confidence ratings are similar and differ maximally about 3% between the ratings with and without ProsAlign. Also the agreement for the existence or non-existence of edge tones (phrase accents and boundary tones merged) is nearly similar to the ones of the human labelers: mean agreement rate with ProsAlign is 86.4% vs. 88.6% mean agreement rate without ProsAlign.

The overall agreement rates between comparisons with ProsAlign on the one side and without ProsAlign on the other are very close together, and show the relatively high quality of the automatically produced labels. The comparison of consistency measurements demonstrate that ProsAlign could be seen as adding just a further human labelers variations in pitch accent and boundary tone placement.

6.4 Discussion

The evaluation of the ProsAlign algorithm showed its performance characteristics: High pitch accents are detected well whereas low pitch accents and especially high boundary tones are detected less reliably. L-L% boundary tones are detected good, although there is still scope for improvement.

	exact PA	PAonly	exact BT	ETonly
LP-L123	50.7	69.5	87.7	85.0
L4-L123	57.3	83.9	91.7	86.0
LP-L234	50.7	69.5	87.7	85.0
L1-L234	60.5	78.6	91.8	88.3
LP-L134	50.6	71.5	90.2	86.4
L2-L134	54.6	74.2	93.1	89.2
LP-L124	60.2	77.2	91.9	89.2
L3-L124	60.6	77.5	94.3	91.1

Table 6.18: Consistency of individual human labelers (L1-L4) with ProsAlign (LP) represented in percent agreement of each labeler with the other three. Edge tones (ET) include phrase accents and boundary tones.

The algorithm's general performance is very close to the ability to detect pitch accents and boundary tones with the same performance of human labelers and is hence a helpful tool for the automatic labeling of large acoustic speech corpora. In view of the variation between human labelers, the algorithms performance may be seen as just another variant of a human labelers transcription. The visual inspection of the automatically set labels shows overall high accuracy (see the examples in appendix A). Pitch accents are labeled at positions that are interpretable as possible positions of pitch accents as well as boundary tone positions.

Regarding the decision whether there is or is not a pitch accent at a specific position in the speech file ProsAlign performs reasonably. However, the algorithm has a tendency to label many more intonational events than a human labeler, that is, it produces a number of insertions. With respect to the placement of boundary tones ProsAlign performs well. There is practically no insertion of boundary tones at positions that are not linguistically motivated. However, the separation of high and low boundary tones appears to be not always reliable, most probably as a result of erroneously calculated F0 values by the pitch tracker.

A number of mismatches in the boundary tone class indicates still problems with respect to the detection of erroneous F0 values at intonation phrase final positions. This point is certainly in need of improvement. A number of improvements to provide more reliable boundary tones detection are imaginable: (1) refinement of the individual weights for features of high and low boundary tones, or (2) inclusion of other features like pre-final lengthening, though this would need the knowledge of the segment boundaries, or (3) usage of a confidence measurement to assign explicit high or low boundary tone labels only in obvious cases and using a unspecified '%' in less clear cases. The latter could be applied similarly for the pitch accents, that is using only "*" for cases which have a low confidence rating and using the explicit accent labels only in cases that exceed a certain confidence thresh-

old. It has to mentioned that an approach like ProsAlign, that solely analyses the course of F0, voicing and RMS without knowledge of the segmental content is in some rare cases not able to make a clear decision about the existence and/or the type of a tone at a specific point in time within the given speech file.

The overall performance qualifies the program as a beneficial automatic prosodic labeler for the linguistic work as well as a research area for studying acoustic features of (phonological) pitch accents or/and modelling the general procedure from the acoustic signal up to the linguistic perception. Poor detection results of certain tone categories also indicate that there could be arguments for a reduction of the number of pitch accents and boundary tones in the underlying phonological model (e.g. H+L*).

ProsAlign did also work on the ToBI corpus in American English, although the bad results for low pitch accents and a number of missing boundary tones indicate that there are either subtle differences in the underlying set of acoustic features or that the set of features has to be adjusted. In the first case a language specific adaptation could help in the second case additional or more refined acoustic features could improve the results. All in all, ProsAlign showed encouraging results for a very diverse set of speech material consisting of utterances from different speakers and languages including diverse speaking styles, unequal recording levels, background noise and cross-talk.

The second evaluation of the reliability of automatically produced GToBI labels showed the high performance of the automatic prosodic aligner. Overall agreement results were reduced only slightly when adding the automatically produced labels to the ones produced by human labelers. The comparison with labels from different human labelers allowed the integration of human variability in prosodic labeling for the evaluation of ProsAlign. ProsAlign shows very similar reliability results as the human labelers.

At what processing stages can errors occur? Looking at the processing levels linearly from the input to the output the following error sources are present: (1) Calculation of F0, RMS, and voicing from the input speech signal. Possible errors are wrong F0 or voicing estimation (cf. section 5.1). RMS values cannot be erroneous, though they do not only include speech but also noise components and are therefore also a possible source of misinterpretations.[4] (2) extraction of features from the course of F0, RMS, and voicing. Here, the wrong estimation of increases or decreases in F0 may occur. Another error source are the number and/or the type of acoustic features chosen. Although means are integrated to account for errors in the F0 values, there remains the possibility that erroneous F0 values mislead the feature extraction. (3) Scoring system: insufficient number of features and incorrectly adjusted relative weighting of scores. (4) Phonological mapping: incomplete or in-

[4] A separation of recordings having a high or a low signal-to-noise ratio may be a direction for future improvements of ProsAlign.

correct rules for deletions, insertions or transformation of tones may significantly influence the output.

Though ProsAlign includes means to detect faulty or microprosodically affected F0 values there remains the possibility that some of them are not captured. However, there is still the possibility in later processing stages to identify them and to apply deletion, insertion, or transformation rules based on contextual information. The latter has to be carefully designed in order to avoid the overcorrection of tones and phonologically unstructured output, which may be caused by sloppy rules.

An important aspect is the status of the phonological verification and deletion rules. Application of such rules enables one to change tone candidates already selected in the scoring module as a result of contextual considerations. For instance, a high pitch accent occurring some 10 milliseconds before a final H-H% boundary tone might need to be deleted on the basis of the knowledge that the boundary tone received a fairly high confidence rating suggesting that the preceding high pitch accent is probably the result of realisation of H%. Derivational as well as durational aspects can be taken in consideration at this stage. Derivational restrictions in the sense of allowable sequences of tones in a given language and durational restrictions based on previously established minimal distance values between tones.

However, in a strict sense the "structuring" influence of the underlying phonological model is provided only by the intonation grammar which defines the legal set of tone sequences for a given language. Durational aspects, like a minimal distance between adjacent tones are partly a consequence of the definition of tones and also a technical consideration within the processing stages of the algorithm in order to prevent tone assignments which are too close to each other. The intonational phonology defines the set of possible intonational events by the limited number of possible pitch accents and boundary tones. Furthermore, the set of legal contours is also defined and could by applied to an unstructured series of tones as calculated by the scoring module. The latter is not fully implemented in ProsAlign since there is the danger of overcorrection. The exact limits of such a selection process have to be established by further studies.

The structuring influence of the phonological model upon the acoustic feature stream may be seen analogously to the structuring influences of the underlying phonological representation in the segmental domain. For instance, a study by the author of this thesis regarding the perception of voicing and vowel length in German has shown that the underlying phonological representation triggers the perception of acoustic stimuli. One and the same vowel was identified differently by human listeners when presented either before a voiced or before a voiceless stop consonant (Braunschweiler, 1994, 1997).

Chapter 7

Conclusions

Prosodic phenomena can be described at several linguistic levels. This thesis dealt with the acoustic as well as the phonological representation of prosody and dealt explicit with the transformation of its acoustic appearance to its phonological description. A central issue of the work are the problems of labeling prosodic phenomena with suitable and linguistically descriptive markers. Three main problem areas may be extracted in this respect: (1) the highly variable speech signal including the problems of reliable extraction of the fundamental frequency, (2) the emormously difficult and time consuming task to label these variable acoustic phenomena satisfactorily according to a given labeling instruction, and (3) the problem that human labelers are often inadequate and unequal when labeling prosodic phenomena. The investigation of these problems lead to the conclusion that an automatic labeling of prosodic phenomena provides a possible solution for them. This thesis described a new method of automatic prosodic annotations. It dealt explicitly with the subject of how to get automatically a discrete phonological description from a phonetically rich signal.

The development of a program for automatic prosodic alignment unfolded a number of challenging issues. To model the complex mapping of discrete phonological entities to acoustic features a new model of processing was developed. This model integrates both bottom-up and top-down operations in order to implement a fully automatic procedure for alignment of prosodic events in speech signals. The model is implemented in a computer program called ProsAlign and has been evaluated on a diverse set of speech material.

In the first part of this thesis a number of intonational phenomena were presented, among which, marking of sentence type, focussing, and disambiguation. Examples from offering contours, calling contours, surprise contours as well as typical contours from declarative and interrogative sentences were presented with accompanying F0 contours. This section showed that one and the same contour can be overlaid on many different sentences, whereas on the other hand also different intonation contours can be put on one and the same text resulting in totally different

interpretations of it. The chapter ended with pointing out typological aspects of intonation and the listing of three grammaticized main usages of intonation across the languages of the world.

Chapter 3 described existing intonation models (IPO, Fujisaki, KIM, RFC-model, Pierrehumbert) and compared two different labeling instructions (ToBI, INTSINT). The phonological description of German intonation was presented in detail followed by the presentation of existing approaches of automatic intonation detection (Pierrehumbert, Wightman & Ostendorf, Taylor, Rapp, Ostendorf & Ross, MOMEL, Verbmobil, ToBI Lite). Several different approaches are used for the automatic recognition of intonational phenomena including rule-based and statistical approaches. One of the main issues appearing through the analysis of these models was the problem of how to differentiate between microprosodically affected F0 movements on the one side and potentially meaningful ones on the other side. This issue is part of the central topic in these approaches, namely which F0 movements are perceptually relevant and which are not. Smoothing the F0 track with more or less effective means before its actual analysis was one of the generally applied strategies to reduce microprosodic effects. Also the detection of segmental content and boundaries was used in order to provide a solution of this challenge. However, more reliable recognition results were often reached only, when the inventory of prosodic labels was drastically reduced to two-way distinctions between accented vs. unaccented syllables or the presence or absence of an intonational phrase boundary. Often results are based on the speech of one speaker only and speech under controlled conditions without changing signal qualities.

Chapter 4 laid out the concept of ProsAlign and dealt with the question: What are the relevant acoustic features for the detection of prosodic events? The experiences during the assessment of acoustic feature criteria were described and statistical results were presented for each individual feature category. The results showed that the chosen feature set was insufficient for a reliable detection of prosodic events. At the end of chapter 4 the phonological mapping procedure in ProsAlign was introduced. This procedure consists of a scoring system that evaluates feature values with regard to their agreement with previously established feature configurations for individual pitch accents. The output of this evaluation is not the final result but further processed by phonological rules that are able to delete, insert, or modify tone candidates. Therefore, a considerable amount of work within the algorithm is provided by the structuring influence of phonological top-down processes. The concept of ProsAlign therefore is based on an analysis of acoustic features on the basis of language independent acoustic relations followed by a scoring procedure that evaluates these acoustic features with respect to their appropriateness for individual phonological pitch accents or boundary tones. Finally the algorithm evaluates the tone candidates provided by the scoring mechanism and applies selection, deletion or insertion rules that produce the final output, that is type and position of pitch accents and boundary tones for the given speech file.

Chapter 5 described the implementation of ProsAlign in a computer program. Here

the individual acoustic features were explained in detail and the procedure to map phonological entities to acoustic feature bundles. As a result of shortcomings in the first feature set additional criteria was integrated. Especially an improved detection of increases and decreases in F0 and with minor importance in RMS was an important step towards more robust selection criteria. The allowance of a limited number of outlying values in F0 increase and decrease estimation proved to be significant. Connected with the latter was the separation of microprosodically affected and faulty F0 values from potentially meaningful ones. Here the additional inspection of the course of voicing and RMS revealed aspects for more selectivity and therefore provides an important contribution towards the solution of these central issues. The phonological module in ProsAlign includes further means to account for these task and allows deletions, insertions, and transformations of tone candidates coming from the acoustic feature evaluation module.

In chapter 6 the ProsAlign program was evaluated in two ways: First the programs output was compared to manually produced labels and second ProsAlign was compared with four human labelers. The first evaluation showed that ProsAlign is able to detect most of the manually annotated tone labels. Comparisons of the automatically transcribed tone labels with two different scenarios (a) when all labels were changed to H* (the category with the highest occurrence) and (b) when labels were placed randomly showed the power of ProsAlign. ProsAlign was evaluated on two corpora, the German GToBI training corpus and the American English ToBI training corpus and showed similar performance results in both cases therefore showing its potential for language independend prosodic annotations. A number of label insertions showed that ProsAlign has a tendency to label slightly more prosodic events as compared to humans. Poor recognition accuracy for specific label categories indicated also that some tone labels may be subject of critical evaluation at the phonological level. The second evaluation, a comparison of ProsAlign's output with the labels annotated by four human labelers measured the inter-transcriber reliability by means already applied to earlier ToBI labeling reliability studies. The results unfolded that ProsAlign performs similarly to the human labelers and showed the quality of the prosodic annotations provided by ProsAlign. The program is a useful tool for automatic labeling of large acoustic speech corpora and for linguistic research. The consistency of the placement of prosodic labels shown by the program is certainly one of its advantages. ProsAlign includes a number of innovative concepts and, last but not least, as a result of its modular architecture it may be adapted to specific purposes.

The next section summarizes the main findings in this thesis. Then possible applications beside the usage as prosodic aligner are addressed. Finally future developments and extensions are discussed.

7.1 Summary of Main Findings

A rule-based approach showed encouraging results by using fairly simple methods as detection criteria of acoustic features combined with the application of a structured phonology - acoustic features mapping. The implementation of the method in a computer program showed solid recognition results. The analysis of inter-transcriber reliability of ProsAlign and four human labelers showed that ProsAlign performs almost similar as another human labeler would. The usage of a score system in between the acoustic features and the phonological entities allowed the integration of information from all input sources and therefore was able to model the vast variability in speech signals. Furthermore, the algorithm proved to be robust against changing input signal qualities and was able to process speech from different speakers and is therefore speaker independent.

Since ProsAlign was also successfully applied to detect pitch accents and boundary tones in other languages than German, it showed its potential for language independent usage. However, a number of decreased recognition results in the American English corpus indicated some caveats here. Therefore, further evaluations have to show whether different languages implement the acoustics of the prosodic labels differently.

The explicit usage of a phonological model of intonational structure in the algorithm showed both positive and negative aspects. Positive was the structuring and therefore complexity reducing influence of the underlying phonological model. Negative was the low selectivity of some categories indicating possible improvements in the phonological model, for instance by reducing the number of possible pitch accents.

With regard to aspects raised during the discussion of the autosegmental-metrical approach of intonational description (see section 3.1.6), it is important to note that ProsAlign does not restrict the position of tones to specific positions like turning points in the F0 contour. Local maxima may correspond to high tones and local minima to low ones. However, there is no necessity for this. Since the ProsAlign model defines a pitch accent as an integrated concept of a bundle of acoustic features and a number of phonological sequence restrictions there is no need to associate tones with turning points. Pitch accents may appear at every F0 value in a voiced stretch of speech and boundary tones may appear at voiced as well as unvoiced positions.

The simultaneous inspection of RMS and voicing proved to be an important addition to the analysis of the F0 course. It improved both the detection of faulty and microprosodically affected F0 values and the detection of pitch accents and boundary tones.

Moreover, the usage of fairly broad acoustic features combined with a subsequent scoring system that evaluates the acoustic features both with regard to their appropriateness for specific pitch accents or boundary tones and possible microprosodi-

cally affected F0 values offers an elegant method to avoid any smoothing or filtering process before the actual analysis takes place.

7.2 Applications for the Program

ProsAlign was first developed as tool for more consistent and faster labeling of large acoustic speech databases. Speech material including (i) diverse recording conditions, (ii) a variable number of speakers, (iii) different languages, and (iv) unrestricted text may be processed automatically and efficiently. Subsequent analysis could first manually verify the automatically set labels or could directly go on to further research.

Another interesting application for the program is in the domain of speech synthesis. The program is currently used for annotating prosodic labels in a unit selection corpus (cf. footnote 16 on page 72) in order to improve the appropriateness of selections. In cases where units consist of the same segmental material and are not distinguished by other means, the existence of prosodic labels can help to differentiate between those. Whenever there is a need for a specific unit with a specific pitch accent or boundary tone, it could be selected from the corpus. Since a manual labeling of pitch accents and boundary tones is very time consuming and might not be consistent enough (see e.g. the critique of manually labeled corpora in the beginning of section 4.2), ProsAlign could significantly improve the set up and the quality of a unit selection synthesis voice.

With regard to automatic speech recognition there can be the following applications of ProsAlign: using the output of the prosody detection module (that works independently of the phonetic segment detectors) as structuring feature for improved speech recognition. Additional information may be extracted from ProsAlign's output and could help to separate competing hypotheses. Alternatively another strategy would be to integrate ProsAlign's results already at preselection stages during the segmental identity recognition. The beneficial usage of prosodic labels in an ASR system has been shown in the Verbmobil project (cf. Nöth et al. 2000 and section 3.3.7), but also that its incorporation, for instance in syntax parsing, is far from being uncontroversial.

ProsAlign could also aid in the detection of incorrect pronunciations in a language learning program (cf. Bagshaw 1994 and the discussion of the limits of this kind of learning aids in Delmonte 2000). The learner could test his freshly acquired language abilities and would receive immediate feedback about the appropriateness of his pronunciation.

7.3 Future

ProsAlign has certainly not yet reached its optimal performance because improvements on several levels are imaginable. The main areas of improvements are:

- the detection of faulty F0 values
- optimization of the individual feature weights in the scoring system
- the detection of boundary tones
- reduction of tone inventory specifically adapted for automatic detection purposes (see e.g. the ToBI Lite approach discussed on page 72)
- task specific adaptations of the algorithm, for instance inclusion of segmental information when it is available as in automatically segmented and manually corrected speech synthesis corpora
- integration of the intonational phonologies of other languages than German and English

The detection of faulty F0 values could be improved by adjusting the individual weights of features within the scoring system. However, since one cannot expect to detect all the faulty or microprosodically affected F0 values, it is also the task of the phonological mapping module to handle these cases. Using contextual information and applying sequence restrictions regarding position and type of tones are possible means for it.

In order to enable the processing of other languages their intonational phonologies have to be incorporated in ProsAlign. In dependence of these individual phonologies, only specific tones are available in the mapping procedure. Tones or tone combinations that do not occur in a given language are not assigned.

It is also imaginable to force a phonologically wellformed output by applying wellformedness restrictions on the sequence of tones. The set of possible tone sequences may be produced by a finite-state grammar. Tone deletions, insertions, transformations and shifts in position might be applied here. However, by doing so, the problem of overcorrection, the change of an acoustically determined tone sequence towards a phonologically well-formed one that does not describe the underlying phonological tone structure of the analyzed utterance, becomes more acute. The precise formulation and the limits of this procedure are something to look at in the future development of ProsAlign. The rules which are applied currently are sequential restrictions that may delete a tone on basis of a neighboring tone that has a higher score and is positionally competing (e.g. is in between 100 ms). What is meant by the enforcement of a phonologically well-formed output is the application of a finite state grammar that allows only tone sequences that fall within the legal set of the grammar.

The integration of speaker individual characteristics could also be a topic for further improvements of ProsAlign. When integrating speaker individual characteristics more reliable tone assignments may be achieved. For instance, the pitch range reflected by the top-line and bottom-line could be integrated in the selection process in order to improve the detection of more detailed label categories like downstepped tones, etc. Speaker individual range values may be handed to the program before processing takes place.

Another expansion of ProsAlign is to enable intermediate outputs at the different processing stages. Results of intermediate processing stages may be of benefit for phonetic research regarding the acoustic characteristics of individual prosodic events. Researchers could use, for instance the output of the acoustic feature extractor as basis for the statistical analysis of acoustic cues of prosodic events. Concentration on specific features (e.g. F0 only) is possible.

Furthermore, the output of confidence values for individual tones could be of benefit for specific purposes. To get an estimation of the reliability of an individual tone it could be accompanied by a confidence value (e.g. when expressing the confidence of an individual tone in percent: H*_85% or L-L%_98%). This could be of benefit in ASR applications or for the manual correction of automatically assigned labels.

Of course, a very important issue in the future development of ProsAlign is the error tracking. By constructing means for error tracking iteratively faulty labels may be eliminated and subsequently improve the algorithms performance. Such means could be a specifically designed label inventory that allows a characterization of the type of error that occurred for a specific label. For instance whether a label is faulty as a result of erroneously interpreted F0 movements or whether a label has to be removed because of the existence of another (more reliable) label within the syllable.

The integration of segmental information could also be thought of. In cases where segmental information (i.e. phoneme, syllable, and word boundaries plus identity; position of lexically stressed syllable) is available, another processing strategy could be implemented in order to include the additional features provided by this information. Durational cues like pre-final lengthening or segment durations could be included in the selection process. Also knowledge about the position of the lexically stressed syllable allows to position labels more exactly according to the theory. The automatic prosody labeling could be adapted to specific purposes, that is for TTS systems there could be first a detection of segmental boundaries. Therefore, positions of stressed syllables are determined and serve as starting point for the tone assignment.

Another aspect concerns the “trainability” of ProsAlign either by manually labeled corpora or by a fully automatic parameter adjustment without prior manual labeling. A criticism of rule-based models is their inability to be automatically trainable (cf. Ostendorf & Ross 1997). It is even imaginable to implement an automatic ad-

justment of the models parameters by building new acoustic feature vector values and adjusting their weights. Ranges and number of acoustic feature values could be automatically estimated from a prosodically annotated speech corpora. However, this would include the drawback that manually established labels might not be consistent enough for training of an automatic recognizer, it could be nevertheless a testable alternative. The acoustic feature detector would acquire the range and concrete size of acoustic feature values relative to the position of the pitch accents and boundary tones provided from the corpus. Building new acoustic feature vector values would mean to estimate minimal and maximal values of each individual feature as well as estimate their relative importance. The latter could be achieved by using predefined weights for the individual features. Additionally, these weights could be checked by using slightly different weights and viewing its impact on recognition accuracy.

Together ProsAlign showed encouraging results for a very diverse set of speech materials consisting of utterances from different speakers and languages including diverse speaking styles, unequal recording levels, background noises and cross-talk. The new model underlying ProsAlign showed its power for this purposes and opens new directions for future research directions.

Appendix A

Examples of Labeled Speech Files

To give the reader a visual impression of the behaviour of ProsAlign the next pages show some randomly selected examples from the GToBI and ToBI corpus as labeled by ProsAlign. Each example illustrates the waveform above the three label files for (1) words, (2) pitch accents and boundary tones as calculated by ProsAlign (illustrated within a frame with thicker lines and including the marker "ProsAlign"), and (3) manually labeled pitch accents and boundary tones; below the label files is the F0 contour. The F0 window shows the pitch range within the given file by the lower base line which is fixed at 60 Hz and a variable upper limit indicated by a dashed line associated with the maximum F0 value next to it. Three series of files are shown: (1) files from the GToBI corpus, (2) files from the ToBI corpus, and finally (3) some examples from various languages. The files represent a good overview of the performance of ProsAlign. Other files are labeled similarly by ProsAlign, some of them deviate more from the manually established comparison files, others are closer or nearly identical to them. The detailed statistical treatment of ProsAlign's performance is presented in chapter 6.

Examples from the GToBI corpus

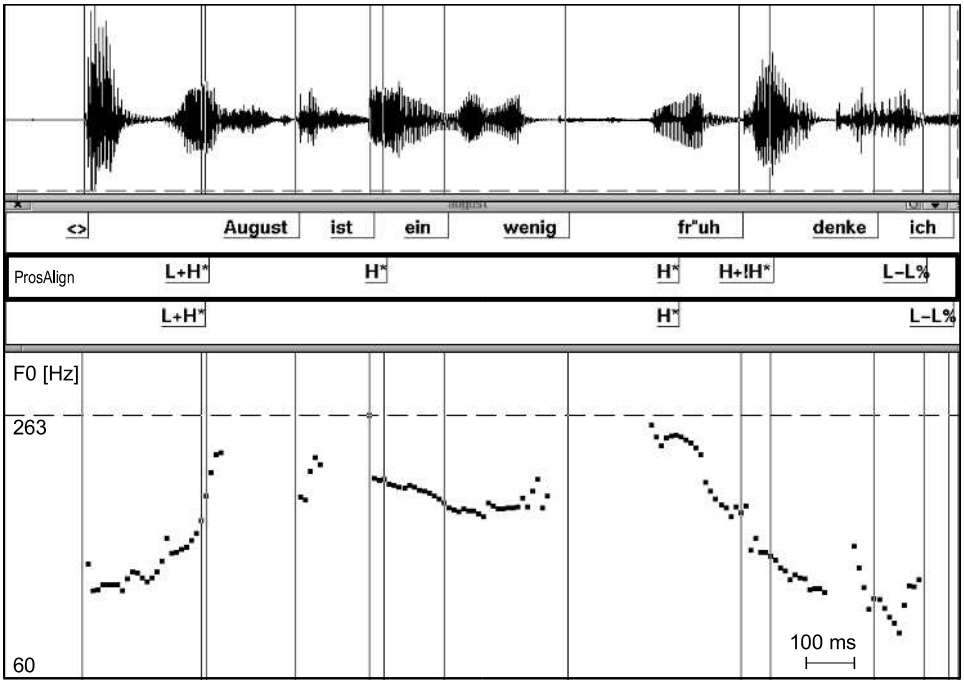


Figure 1: File "august" from GToBI corpus as labeled by ProsAlign and manually. The text is: August ist ein wenig früh denke ich. *August is a little bit early I think.* The two pitch accents in the manual label file are perfectly detected by ProsAlign in position and type but two additional tones are inserted. The final boundary tone is annotated slightly too early before the fricative of "ich" [ɪx] *I* has ended.

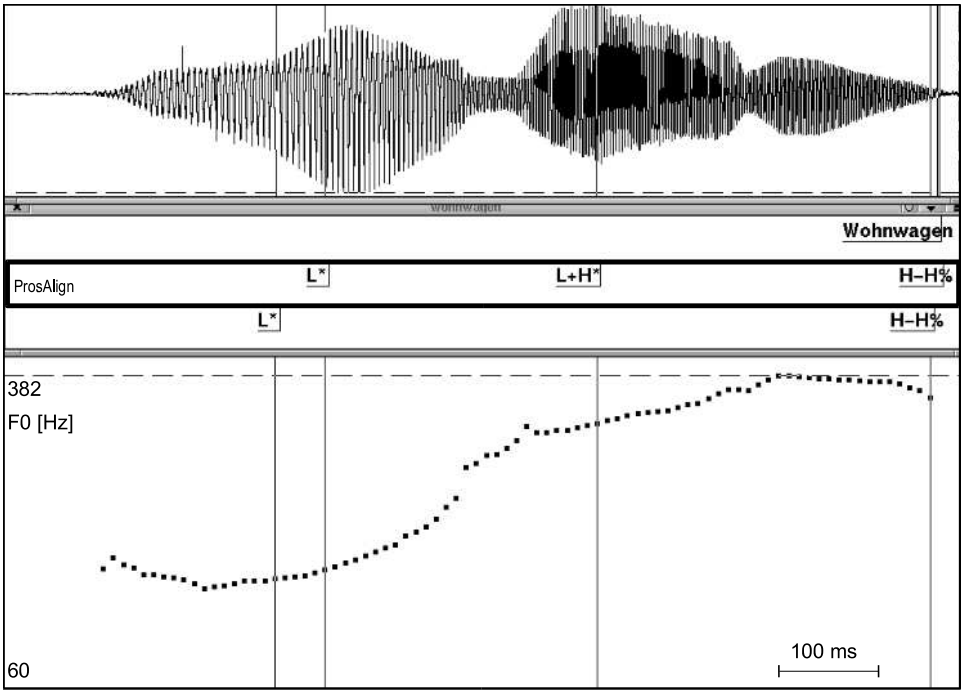


Figure 2: File "wohnwagen" from GToBI corpus as labeled by ProsAlign and manually. The text is: "Wohnwagen?" *camper?* Both the low pitch accent and the high boundary tone are detected by ProsAlign but the rise in the second syllable was interpreted as a L+H* pitch accent by ProsAlign.

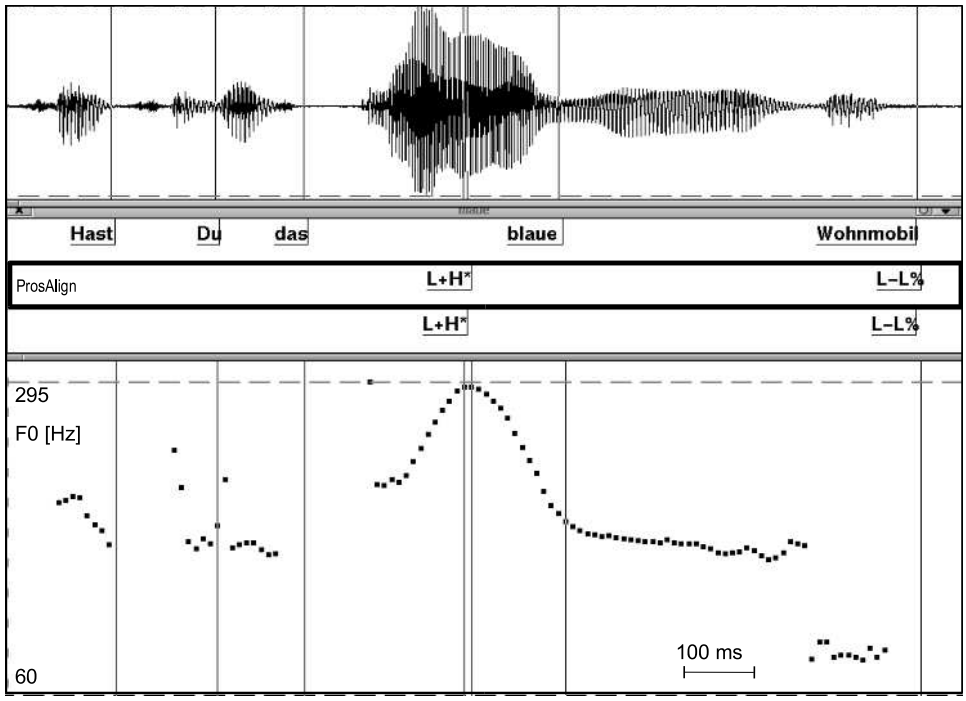


Figure 3: File "blaue" from GToBI corpus as labeled by ProsAlign and manually. The text is: "Hast du das blaue Wohnmobil? *Do you have the blue camper?* This example shows a perfect match of the manually annotated labels and those transcribed by ProsAlign.

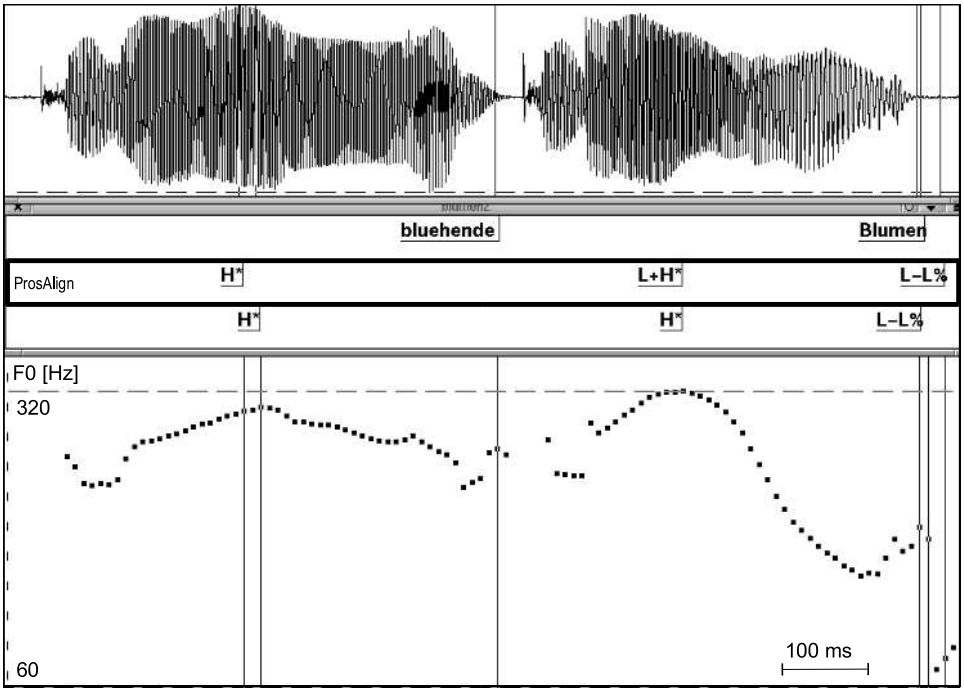


Figure 4: File "blumen2" from GToBI corpus as labeled by ProsAlign and manually. The text is: "blühende Blumen" *blooming flowers*. ProsAlign detected all the manually annotated tones but decided to label the second pitch accent as L+H* instead of H* as in the manual label file. The final boundary tone is labelled slightly too late.

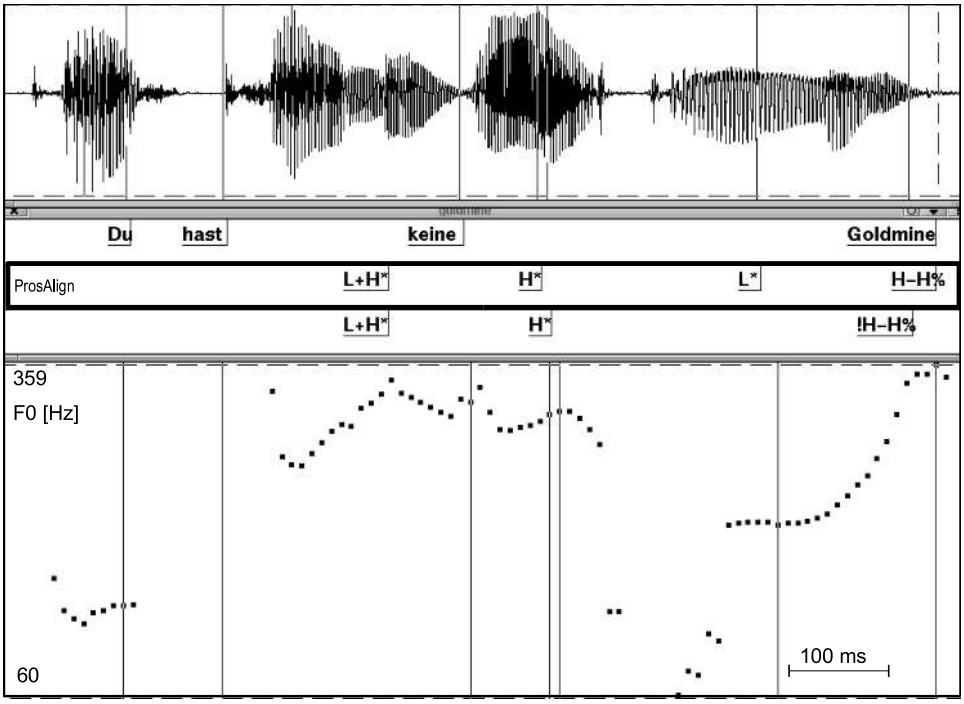


Figure 5: File "goldmine" from GToBI corpus as labeled by ProsAlign and manually. The text is: "Du hast keine Goldmine?" *You do not have a gold mine?* The two manually labeled pitch pitch accents are also detected by ProsAlign. The final rise is correctly recognized with a H-H% by ProsAlign but before there is a L* pitch accent inserted.

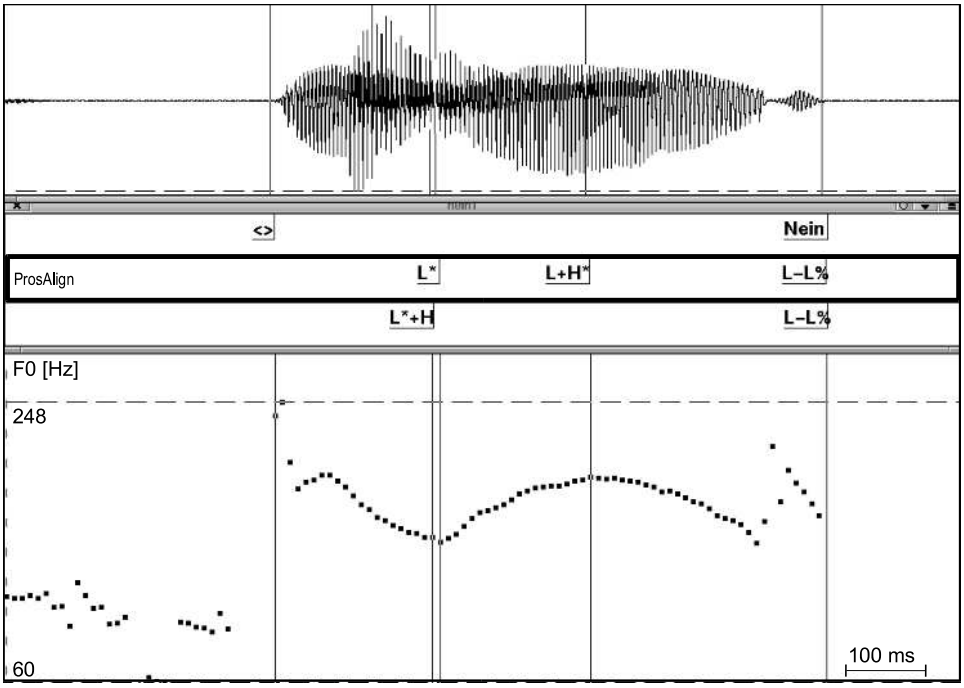


Figure 6: File "nein" from GToBI corpus as labeled by ProsAlign and manually. The text is: "Nein" *No*. The low-high-low movement is described by ProsAlign with two pitch accents and a final low boundary tone. The human labeler annotates only one pitch accent and one final boundary tone.

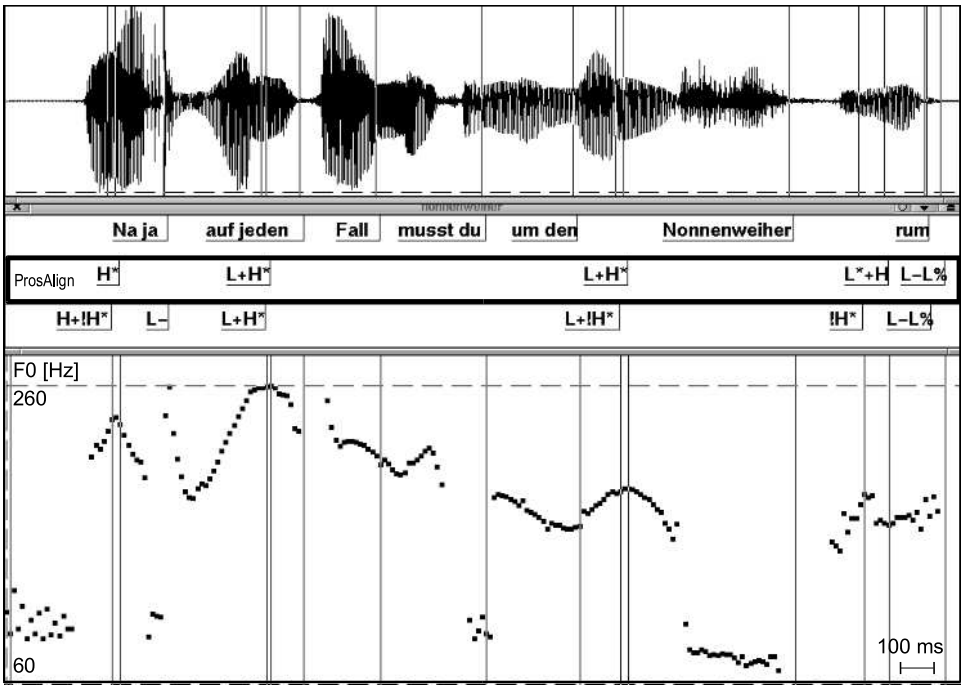


Figure 7: File "nonnenweiher" from GToBI corpus as labeled by ProsAlign and manually. The text is: "Naja, auf jeden Fall musst du um den Nonnenweiher rum." *Well, at all events you have to go around the Nonnenweiher*. ProsAlign recognized the pitch accent on the first word as H* whereas the human labeler labeled a H+!H* at this position. The low intermediate boundary tone was not detected by ProsAlign. The F0 movements at the end are marked as L*+H movement by ProsAlign whereas the manual label file shows a !H*. Though the F0 values at the end are rising ProsAlign is not mislead but indicates a low final boundary tone as the human labeler did also.

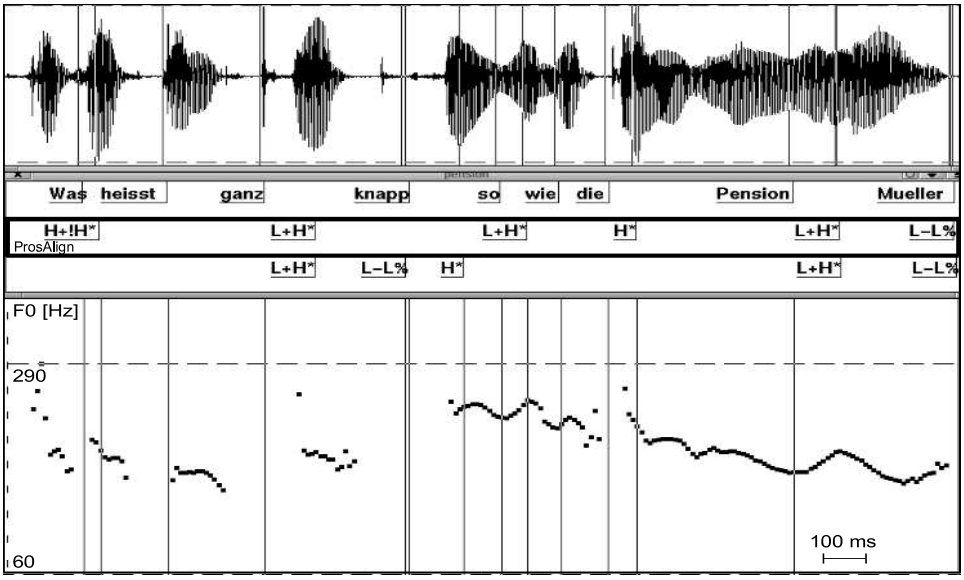


Figure 8: File "pension" from GToBI corpus as labeled by ProsAlign and manually. The text is: "Was heisst ganz knapp, so wie die Pension Müller" *What means just sufficient, the same as the guest-house Müller.* ProsAlign inserts an H+!H* pitch accent at the second word and does not recognize the low boundary tone after "knapp" *just sufficient* probably as a result of the non-existence of a pause afterwards. In the second intonation phrase ProsAlign did not label an H* accent on the first word but labled an L+H* on the second one. ProsAlign also labels "Pension" *guest-house* as accented.

Examples from the ToBI corpus

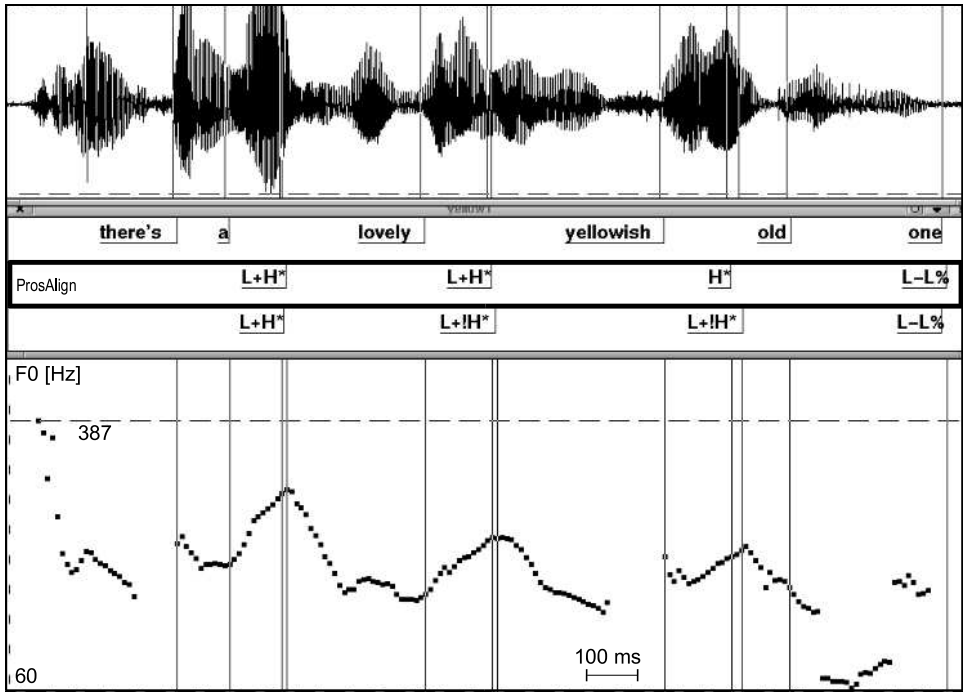


Figure 9: File "yellow1" from ToBI corpus as labeled by ProsAlign and manually. Fairly good coincidence of manually and automatically established labels.

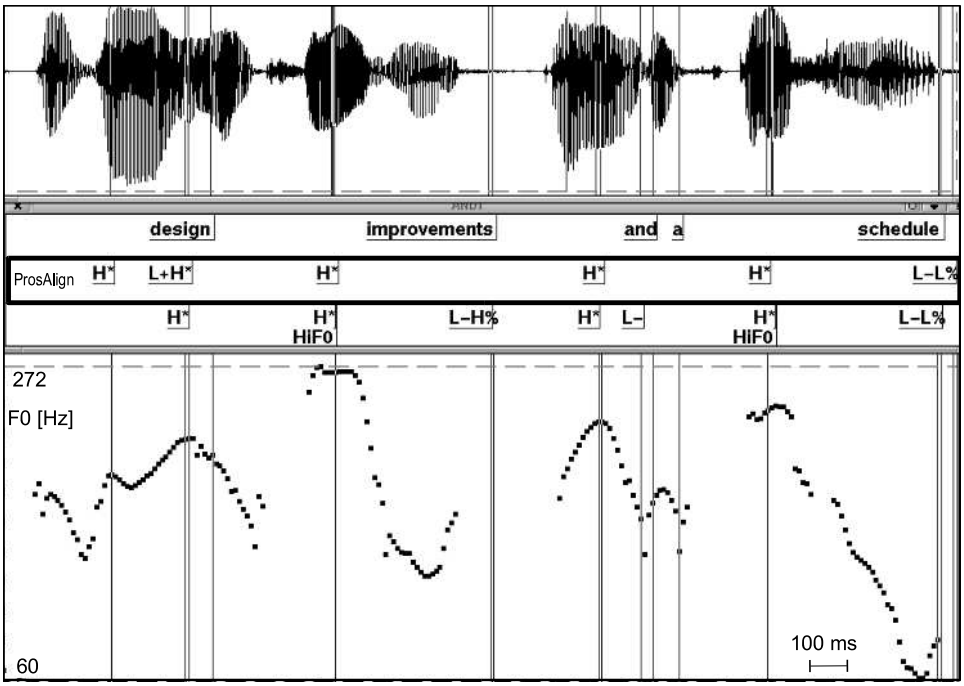


Figure 10: File "AND1" from ToBI corpus as labeled by ProsAlign and manually. ProsAlign inserts a H* pitch accent on "design" and does not recognize the L-H% boundary tone after "improvements", also the intermediate phrase boundary tone after "and" is not annotated by ProsAlign, though the author would not label one at this point.

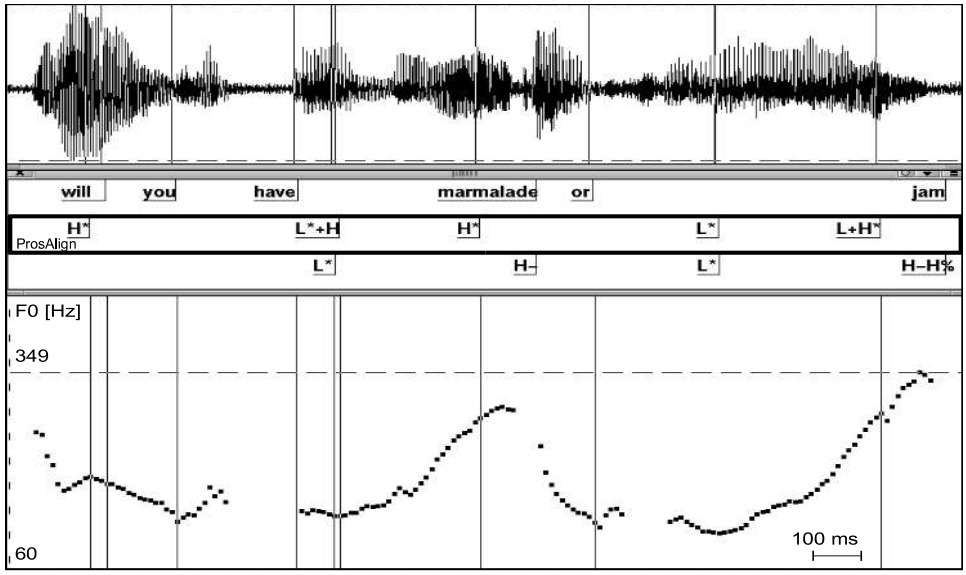


Figure 11: File "jam1" from ToBI corpus as labeled by ProsAlign and manually. ProsAlign interprets the rise in "will" as H* pitch accent and puts a L*+H instead of a L* on "marmalade" The intermediate phrase boundary tone after "marmalade" is not recognized instead the rise is labeled by a H* pitch accent by ProsAlign. The low-high movement on "jam" is labeled as L* followed by L+H* instead of the final high boundary tone; which was probably not detected as a result of the missing right context.

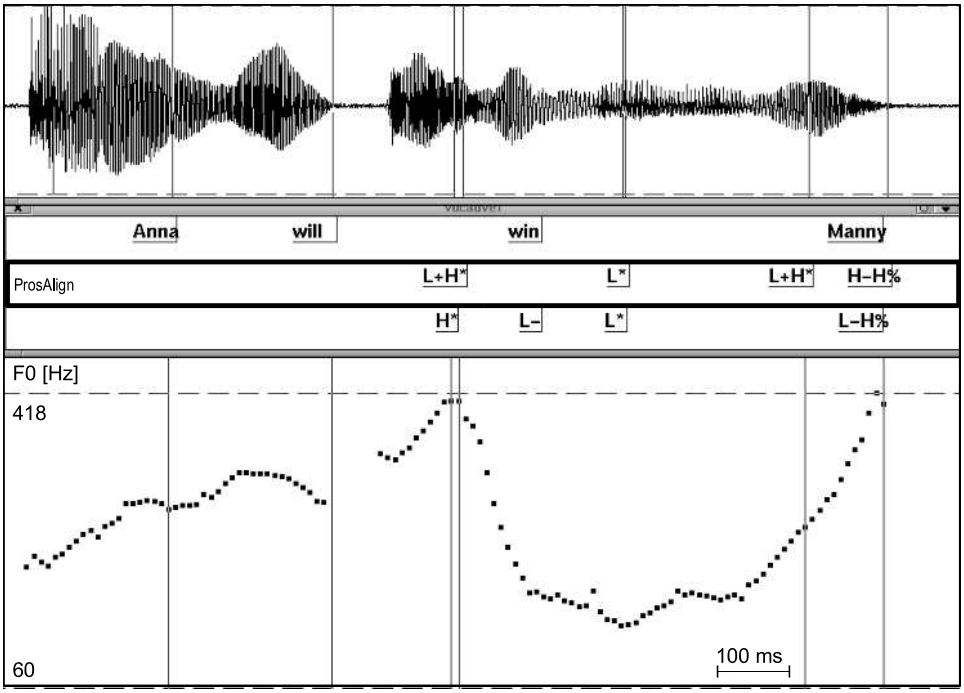


Figure 12: File "vocative1" from ToBI corpus as labeled by ProsAlign and manually. Here ProsAlign does once again not label the intermediate low boundary tone and annotates the final rise in "Manny" as L+H* followed by H-H% insred of the L* and L-H% series in the manual label file.

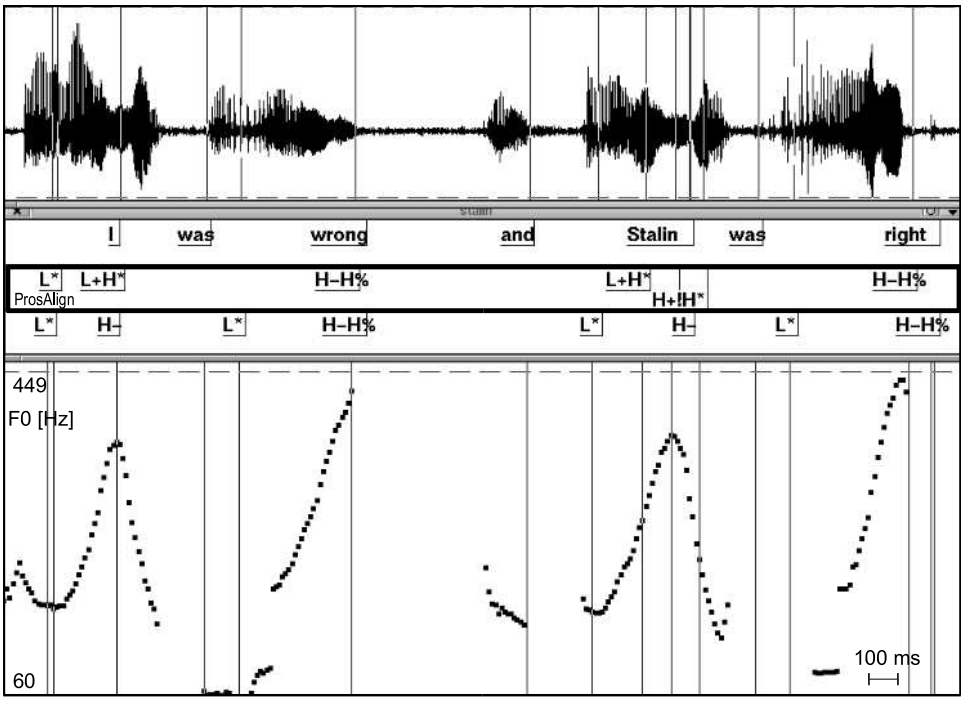


Figure 13: File "stalin" from ToBI corpus as labeled by ProsAlign and manually. ProsAlign labeled the rise in "I" as an L+H* accent whereas the human labeler placed a H- intermediate phrase boundary there. The L* on "wrong" was not detected by ProsAlign, but the high boundary tone afterwards perfectly. The rising movement in "Stalin" was labeled as L+H* within the rise and with a invisible H* label (covered by the H+!H* label) at the maximum by ProsAlign whereas the manual label file shows a L* here and a high intermediate boundary tone at the end. The L* pitch accent on "right" was not detected by ProsAlign but the final high boundary tone perfectly in type but slightly too early.

Examples from various languages

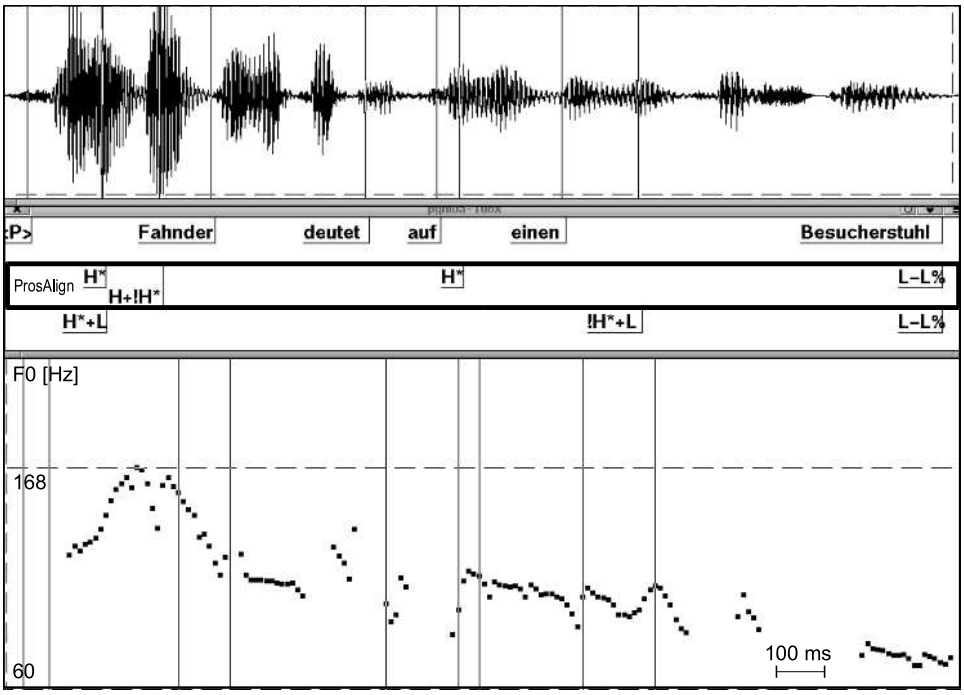


Figure 14: Example of Northern Standard German (this example is also described in Fitzpatrick-Cole, 1999, p. 943) as labeled by ProsAlign and manually. The text is: Fahnder deutet auf einen Besucherstuhl. *The detective points to a visitor's chair.* ProsAlign labeled the rising movement in the first syllable of "Fahnder" *detective* as H* whereas the human labeler inserted a H*+L pitch accent. The second syllable of Fahnder was labeled incorrectly with an H+!H* pitch accent by ProsAlign. Furthermore, ProsAlign inserted a H* on "einen" *a* and did not transcribe the accent on "Besucherstuhl" *visitor's chair.* Cf. the Bern Swiss German version in figure 15.

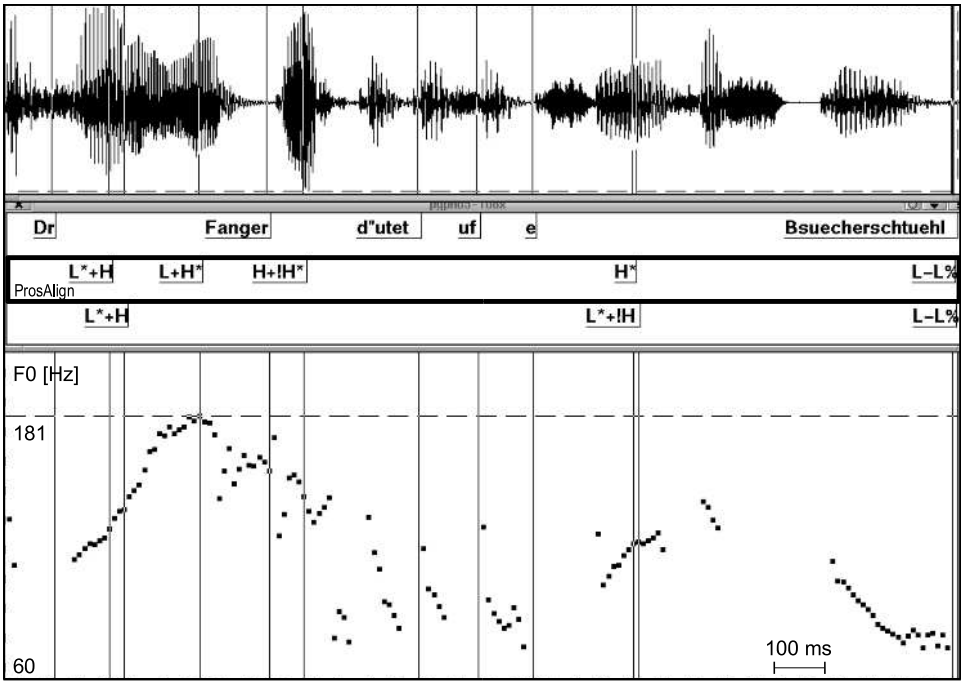


Figure 15: This is the same text as in figure 14 above but in a Bern Swiss German version. The text is: "Dr Fanger dütet uf e Bsuecherschtuhl." *The detective points to a visitor's chair.* The characteristic L*+H movement in Bern Swiss German (cf. Fitzpatrick-Cole, 1999, p. 943) is recognized by ProsAlign but about 140 ms later ProsAlign labels a L+H* at the maximum of the F0 rise. Furthermore ProsAlign labels "dütet" with an H+!H* and instead of the manually as L*+H labeled "Bsuecherschtuhl" ProsAlign recognizes this accent as H*.

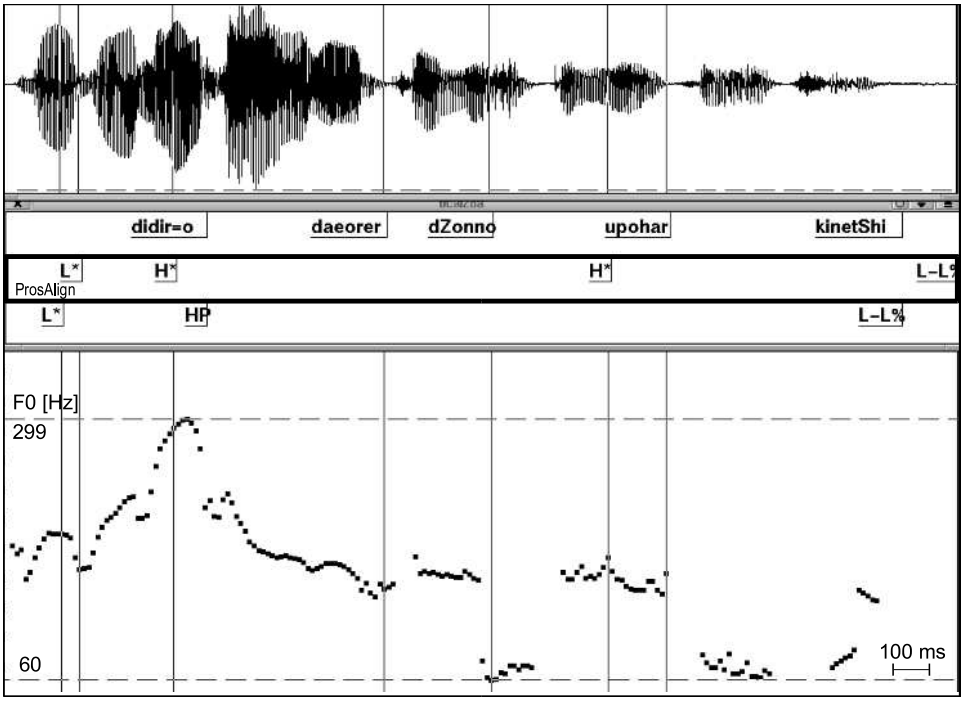


Figure 16: Example from Bengali as labeled by ProsAlign and manually. In this sentence (see linguistic transcription below) the first word is focussed which is obligatorily signalled in Bengali by the contour L*+H$_P$ (see Lahiri & Fitzpatrick-Cole 1999). The initial L* labeled by ProsAlign is slightly too late and probably a result of the high-low-high movement at this place. The HP which is the high boundary tone of the first intermediate phrase is not recognized by ProsAlign as boundary tone but as H* pitch accent. The H* inserted on "upohar" sounds legal to the author. L-L% is labeled too late.

didi-ro	dæor-er	dʒonno	upohar	kinetʃʰi
elder	husband's	for	present	buy-PERF-PRES-1P
sister-GEN=also	younger			
	brother-GEN			

*I bought a present for **sister's** brother-in-law, too.*

Appendix B

Notes on the Computer Implementation

In the beginning the program consisted of several awk and sed programs combined in a shell script. Later it was transferred into C/C++ and now consists of about 12800 lines of code. ProsAlign processes an ESPS/waves input file containing the parameters F0, RMS and voicing extracted from a speech file of 8.8 seconds duration and sampled at 16000 Hz in about 1.5 seconds on a Linux machine with a 700 MHz processor and 384 MB RAM. This is certainly fast enough for most of the applications ProsAlign is intended to be used for. Usually ProsAlign is used in a batch process in order to label a series of files of a more or less large acoustic speech corpus.

ProsAlign was originally developed with the ESPS/waves environment and therefore it takes an ESPS/waves F0 file as input and computes the positions and types of pitch accents and boundary tones in an ESPS/waves label file. However, the general procedure is not bound to this kind of environment but may also be adapted to other file formats and may also be used without this program package.

Bibliography

Adriaens, L. M. H. (1991). *Ein Modell deutscher Intonation*. PhD thesis, Technische Universiteit Eindhoven.

Arvaniti, A. & Baltazani, M. (to appear). Intonational analysis and prosodic annotation of Greek spoken corpora. In S.-A. Jun (Ed.), *Prosodic Models and Transcription: Towards Prosodic Typology*. Oxford University Press.

Bagshaw, P. C. (1994). *Automatic prosodic analysis for computer aided pronunciation teaching*. PhD thesis, University of Edinburgh.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., & Niemann, H. (1999). Prosodic feature evaluation: Brute force or well designed? In *Proceedings of the 14th International Conference of Phonetic Sciences (ICPhS)*, volume 3 (pp. 2315–2318). San Francisco, CA.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., & Niemann, H. (2001a). Boiling down prosody for the classification of boundaries and accents in German and English. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)* (pp. 2781–2784). Aalborg, Denmark.

Batliner, A., Buckow, J., Niemann, H., Nöth, E., & Warnke, V. (2000). The prosody module. In W. Wahlster (Ed.), *VERBMOBIL: Foundations of Speech-to-Speech Translations* (pp. 106–121). New York.

Batliner, A., Kompe, R., Kießling, A., Mast, M., Niemann, H., & Nöth, E. (1998). M = syntax+prosody: A syntactic-prosodic labeling scheme for large spontaneous speech databases. *Speech Communication*, 25, 193–222.

Batliner, A., Kompe, R., Kießling, A., Niemann, H., & Nöth, E. (1996). Syntactic-prosodic labeling of large spontaneous speech data-bases. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)* (pp. 1720–1723). Philiadelphia, PA.

Batliner, A., Nöth, E., Buckow, J., Huber, R., Warnke, V., & Niemann, H. (2001b). Whence and whither prosody in automatic speech understanding: A case study. In M. Bacchiani, J. Hirschberg, D. Litman, & M. Ostendorf (Eds.), *Proceedings of the Workshop on Prosody and Speech Recognition* (pp. 3–12). Red Bank, NJ.

Baumann, S., Grice, M., & Benzmüller, R. (2001). GToBI - a phonological system for the transcription of German intonation. In S. Puppel & G. Demenko (Eds.), *Prosody 2000. Speech Recognition and Synthesis* (pp. 21–28). Poznan: Adam Mickiewicz University, Faculty of Modern Languages and Literature.

Beckman, M. E. & Ayers, G. M. (1997). *Guidelines for ToBI Labelling (version 3, March 1997)*. Technical report, Ohio State University.

Beckman, M. E. & Jun, S.-A. (1996). *K-ToBI (KOREAN ToBI) Labeling Conventions (version 2.1, revised November 1996)*. Technical report, Ohio State University.

Beckman, M. E. & Pierrehumbert, J. B. (1986). Intonational structure in English and Japanese. *Phonology Yearbook*, 3, 255–310.

Bierwisch, M. (1966). Regeln für die Intonation deutscher Sätze. In *Untersuchungen über Akzent und Intonation im Deutschen*, volume 7 of *Studia Grammatica* (pp. 99–201). Berlin: Akademie-Verlag.

Bolinger, D. (1972). Accent is predictable (if you're a mind reader). *Language*, 48, 633–644.

Bolinger, D. (1978). Intonation across languages. In J. Greenberg (Ed.), *Universals of Human Language*, volume 2 (pp. 471–524). Palo Alto, CA: Stanford University Press.

Bolinger, D. (1986). *Intonation and its Parts*. Palo Alto, CA: Stanford University Press.

Bolinger, D. (1989). *Intonation and its Uses*. Palo Alto, CA: Stanford University Press.

Braunschweiler, N. (1994). *Stimmhaftigkeit und Vokallänge im gesprochenen Deutsch: Produktions- und Perzeptionsexperimente zur Bestimmung der akustischen Schlüsselparameter*. Master thesis, University of Konstanz, Germany.

Braunschweiler, N. (1997). Integrated cues of voicing and vowel length in German: A production study. *Language and Speech*, 40, 353–376.

Bruce, G. (1977). Swedish word accents in sentence perspectives. In *Developing the Swedish intonation model. Working papers, Department of Linguistics and Phonetics, University of Lund*, volume 22 (pp. 51–116). Lund: Gleerup.

Buckow, J., Batliner, A., Huber, R., Niemann, H., & Nöth, E. (2000). Detection of prosodic events using acoustic-prosodic features and part-of-speech tags. In *Proceedings of the International Workshop Speech and Computer (SPECOM'00)* (pp. 63–66). St. Petersburg.

Butzberger, J., Ostendorf, M., Price, P., & Shattuck-Hufnagel, S. (1990). Isolated word intonation recognition using hidden Markov models. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 773–776). Albuquerque, NM.

Campbell, N. (1996). Autolabelling Japanese ToBI. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)* (pp. 2399–2402). Philadelphia, PA.

Campbell, W. (1994). Combining the use of duration and F0 in an automatic analysis of dialogue prosody. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 3 (pp. 1111–1114). Yokohama, Japan.

Campione, E., . V. J. (2001). Semi-automatic tagging of intonation in French spoken corpora. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 90–99). Lancaster, U.K.: Lancaster University, UCREL.

Campione, E., Hirts, D., & Véronis, J. (2000). Automatic stylisation and modelling of French and Italian intonation. In A. Botinis (Ed.), *Intonation: Analysis, Modelling and Technology*, volume 15 of *Text, Speech and Language Technology* (pp. 185–208). Dordrecht: Kluwer.

Chen, F. & Withgott, M. (1992). The use of emphasis to automatically summarize a spoken discourse. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 229–232). San Francisco, CA.

Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row.

Clements, G. N. & Hume, E. V. (1995). The internal organization of speech sounds. In J. A. Goldsmith (Ed.), *The Handbook of Phonology* (pp. 245–306). Cambridge, MA: Blackwell.

Comrie, B. (1984). Russian. In W. S. Chisholm, L. T. Milic, & J. Greppin (Eds.), *Interrogativity*, number 4 in Typological Studies in Language (pp. 7–46). Amsterdam: John Benjamins.

Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.

Crystal, D. (1995). *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press.

Delmonte, R. (2000). SLIM prosodic automatic tools for self-learning instruction. *Speech Communication*, 30, 145–166.

ESPS/waves+ (2001). *Manuals of Product Release 5.3*. Entropic, Inc., Washington, DC.

Fitzpatrick-Cole, J. (1999). The alpine intonation of Bern Swiss German. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS)* (pp. 941–944). San Francisco.

Fujisaki, H. (1997). Prosody, models, and spontaneous speech. In Y. Sagisaki, N. Campbell, & N. Higuchi (Eds.), *Computing Prosody* (pp. 27–42). New York: Springer.

Fujisaki, H. (1981). Dynamic characteristics of voice fundamental frequency in speech and singing – Acoustical analysis and physiological interpretations. In *Proceedings of the 4th F.A.S.E Symposium on Acoustics and Speech*, volume 2 (pp. 57–70).

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In P. F. Mac-Neilage (Ed.), *The Production of Speech* (pp. 39–55). Berlin: Springer.

Fujisaki, H. & Hirose, K. (1982). Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. *Preprints of papers for the working group on intonation, 13th International Congress of Linguists*, (pp. 57–70). Tokyo.

Féry, C. (1993). *German intonational patterns*. Linguistische Arbeiten, 285. Tübingen: Niemeyer.

Gårding, E. (1977). The importance of turning points for the pitch patterns of Swedish accents. In L. M. Hyman (Ed.), *Studies in Stress and Accent*, volume 4 of *Southern California Occasional Papers in Linguistics* (pp. 27–35). Los Angeles, CA: Department of Linguistics, University of Southern California.

Geoffrois, E. (1993). A pitch contour analysis guided by prosodic event detection. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)* (pp. 793–796). Berlin, Germany.

Gibbon, D. (1976). *Perspectives of Intonation Analysis*. Bern: Lang.

Gibbon, D. (1998). Intonation in German. In D. Hirst & A. Di Christo (Eds.), *Intonation Systems: A Survey of Twenty Languages* (pp. 78–95). Cambridge: Cambridge University Press.

Goldsmith, J. (1976). *Autosegmental Phonology*. PhD thesis, MIT, Distributed by IULC and published 1979 by Garland Press, New York.

Grice, M. (1995). *The intonation of interrogation in Palermo Italian: Implications for intonation theory*. Tübingen: Niemeyer.

Grice, M., Baumann, S., & Benzmüller, R. (to appear). German intonation in autosegmental-metrical phonology. In S.-A. Jun (Ed.), *Prosodic Typology and Transcription: A Unified Approach*, Prosodic Typology. Oxford University Press.

Grice, M. & Benzmüller, R. (1995). Transcription of German intonation using ToBI-tones: The Saarbrücken System. *PHONUS (Research Report, Institute of Phonetics, University of the Saarland) 1*, (pp. 33–51).

Grice, M. & Benzmüller, R. (1997). Trainingsmaterialien zur Etikettierung deutscher Intonation mit GToBI and accompanying speech materials. *Phonus 3, Institute of Phonetics, University of the Saarland*, (pp. 9–34).

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J., & Batliner, A. (1996). Consistency in transcription and labelling of German intonation with GToBI. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, volume 3 (pp. 1716–1719). Philadelphia, PA.

Gussenhoven, C. (1984). *On the Grammar and Semantics of Sentence Accents*. Dordrecht: Foris.

Gussenhoven, C. (2002). Intonation and interpretation: Phonetics and phonology. In *Proceedings of the 1st International Conference on Speech Prosody* (pp. 47–57). Aix on Provence, France.

Gussenhoven, C. & Jacobs, H. (1998). *Understanding Phonology*. London: Arnold.

Gussenhoven, C., Rietveld, T., & Terken, J. (1999). ToDI - transcription of Dutch intonation. *http://lands.let.kun.nl/todi/todi/home.htm*.

Haggard, M., Summerfield, Q., & Roberts, M. (1981). Psychoacoustical and cultural determinants of phoneme boundaries: evidence from trading F0 cues in the voiced-voiceless distinction. *Journal of Phonetics*, 9, 49–62.

Halliday, M. A. K. (1967). *Intonation and Grammar in British English*. The Hague: Mouton.

Hayes, B. & Lahiri, A. (1991). Bengali intonational phonology. *Natural Language and Linguistic Theory*, 9, 47–96.

Helfrich, H. (1985). *Satzmelodie und Sprachwahrnehmung: psycholinguistische Untersuchungen zur Grundfrequenz*. Berlin: de Gruyter.

Hess, W. (1983). *Pitch determination of speech signals: Algorithms and Devices*. Berlin: Springer.

Hess, W., Batliner, A., Kiessling, A., Kompe, R., Nöth, E., Petzold, A., Reyelt, M., & Strom, V. (1997). Prosodic modules for speech recognition and understanding in VERBMOBIL. In Y. Sagisaki, N. Campbell, & N. Higuchi (Eds.), *Computing Prosody* (pp. 361–382). New York: Springer.

Hirai, T., Higuchi, N., & Saigisaka, Y. (1995). A study of a scale for automatic prediction of prosodic phrase boundary based on the distribution of parameters from a critical damping model. In *Proceedings Spring Meeting, Acoustics Society Japan* (pp. 315–316). (in Japanese).

Hirst, D. & Di Christo, A. (1998). *Intonation Systems: A Survey of Twenty Languages*. Cambridge, UK: Cambridge University Press.

Hirst, D., Di Christo, A., & Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. *Internet paper. http://194.57.187.30/∼hirst/articles/2000 Hirst&al.pdf*, (pp. 1–21).

Hirst, D. J. (1980). Un modéle de production de l'intonation. *Traveaux de l'Institut de Phonétique d'Aix*, (pp. 297–315).

Hirst, D. J. (1983). Structures and categories in prosodic representations. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and Measurements* (pp. 93–109). Berlin: Springer.

Hirst, D. J. & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Traveaux de l'Institut de Phonétique d'Aix-en-Provence*, 15, 75–85.

Huber, D. (1988). Laryngealization as a boundary cue in read speech. *Working Papers*, 34, 66–67. Lund.

Hunt, A. J. & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (ICASSP)*, volume 1 (pp. 373–376). Atlanta, Georgia.

Jensen, U., Moore, R., Dalsgaard, P., & Lindberg, B. (1993). Modelling of intonation contours at the sentence level using CHMMs and the 1961 O'Connor and Arnold scheme. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)* (pp. 785–788). Berlin, Germany.

Kamp, H. & Reyle, U. (1993). *From Discourse to Logic*. Dordrecht: Kluwer Academic Publishers.

Keijsper, C. E. (1983). Comparing Dutch and Russian pitch contours. *Russsian Linguistics*, 7, 101–154.

Kießling, A. (1997). *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Aachen, Germany: Shaker.

Kiparsky, P. (1966). Über den deutschen Akzent. In *Untersuchungen über Akzent und Intonation im Deutschen*, volume 7 of *Studia Grammatica* (pp. 69–98). Berlin: Akademie-Verlag.

Kohler, K. J. (1991). *Studies in German Intonation*, volume 25 of *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung*. Universität Kiel.

Kohler, K. J. (1997). Modelling prosody in spontaneous speech. In Y. Sagisaki, N. Campbell, & N. Higuchi (Eds.), *Computing Prosody* (pp. 187–210). New York: Springer.

Kompe, R. (1997). *Prosody in Speech Understanding Systems*. Lecture Notes in Artificial Intelligence, 1307. Berlin: Springer.

Kompe, R., Batliner, A., Kiessling, A., Kilian, U., Niemann, H. H., & Nöth, E. (1994). Automatic classification of prosodically marked phrase boundaries in German. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 173–176). Adelaide, Australia.

Kompe, R., Kiessling, A., Niemann, H., Nöth, E., Schukat-Talamazzini, E., Zottman, A., & Batliner, A. (1995). Prosodic scoring of word hypotheses graphs. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)* (pp. 1333–1336). Madrid, Spain.

Kutik, E. J., Cooper, W. E., & Boyce, S. (1983). Declination of fundamental frequency in speaker's production of parenthetical and main clauses. *Journal of the Acoustical Society of America*, 73, 1731–1738.

Ladd, D. R. (1978). Stylized intonation. *Language*, 59, 721–759.

Ladd, D. R. (1980). *The structure of intonational meaning: evidence from English*. Bloomington: Indiana University Press.

Ladd, D. R. (1983). Levels versus configurations, revisited. In F. B. Agard, G. B. Kelly, A. Makkai, & V. B. Makkai (Eds.), *Essays in honour of Charles F. Hockett* (pp. 93–131). Leiden: Brill.

Ladd, R. D. (1988). Declination 'reset' and the hierarchical organization of utterances. *Journal of the Acoustical Society of America*, 84, 530–544.

Ladd, R. D. (1996). *Intonational Phonology*, volume 79 of *Cambridge Studies in Linguistics*. Cambridge, UK: Cambridge University Press.

Lahiri, A. & Fitzpatrick-Cole, J. (1999). Emphatic clitics and focus intonation in Bengali. In R. Kager & W. Zonneveld (Eds.), *Phrasal Phonology* (pp. 119–144). Nijmegen: Nijmegen University Press.

Laver, J. (1994). *Principles of Phonetics*. Cambridge textbooks in linguistics. Cambridge University Press.

Lea, W. (1980). Prosodic aids to speech recognition. In W. Lea (Ed.), *Trends in Speech Recognition* (pp. 166–205). Englewood Cliffs, NJ: Prentice-Hall.

Leben, W. (1973). *Suprasegmental Phonology*. PhD thesis, MIT.

Leben, W. (1976). The tones in English intonation. *Linguistic Analysis*, 2, 69–107.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Li, C. N. & Thompson, S. A. (1984). Mandarin. In W. Jr. Chisholm (Ed.), *Interrogativity*, volume 4 of *Typological Studies in Language (TSL)* (pp. 47–61). Amsterdam/Philadelphia: John Benjamins.

Liberman, M. (1975). *The Intonational System of English*. PhD thesis, MIT, New York: Garland Press.

Liberman, M. & Pierrehumbert, J. B. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. Oerhle (Eds.), *Language Sound Structure* (pp. 157–233). Cambridge, MA: MIT Press.

Liberman, M. & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, 8, 249–336.

Ljolje, A. & Fallside, F. (1987). Recognition of isolated prosodic patterns using hidden markov models. *Computer Speech and Language*, 2, 27–33.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman and Co.

Mayer, J. (1995). Transcribing German intonation - the Stuttgart system. *http://www.ims.uni-stuttgart.de/phonetik/joerg/labman/STGTsystem.html*.

McCarthy, J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, 17, 207–263.

Mixdorff, H. (2000). A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3 (pp. 1281–1284). Istanbul, Turkey.

Möbius, B. (1993). *Ein quantitatives Model der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläufen*. Tübingen: Niemeyer.

Möbius, B., Pätzold, M., & Hess, W. (1993). Analysis and synthesis of F0 contours by means of Fujisaki's model. *Speech Communication*, 13, 53–61.

Nakai, M., Singer, H., Sagisaka, Y., & Shimodaira, H. (1995). Automatic prosodic segmentation by F0 clustering using superposition modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 624–627). Detroit, Michigan.

Nakai, M., Singer, H., Sagisaki, Y., & Shimodaira, H. (1997). Accent phrase segmentation by F0 clustering using superpositional modelling. In Y. Sagisaki, N. Campbell, & N. Higuchi (Eds.), *Computing Prosody: Computational Models for Processing Spontaneous Speech* chapter 22, (pp. 343–359). New York.

Nespor, M. & Vogel, I. (1986). *Prosodic Phonology*. Dordrecht: Foris.

Niemann, H., Nöth, E., Kiessling, A., Kompe, R., & Batliner, A. (1997). Prosodic processing and its use in VERBMOBIL. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2 (pp. 77–78). Munich, Germany.

Noguchi, H., Kiriyama, K., Matsuda, H., Taniguchi, M., Den, Y., & Katagiri, Y. (1999). Automatic labeling of Japanese prosody using J-ToBI style description. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 5 (pp. 2259–2262). Budapest, Hungary.

Nöth, E. (1991). *Prosodische Information in der automatischen Spracherkennung - Berechnung und Anwendung*. Tübingen, Germany: Niemeyer.

Nöth, E., Batliner, A., Kießling, A., Kompe, R., & Niemann, H. (2000). VERBMOBIL: The use of prosody in the linguistic components of a speech understanding system. In *IEEE Transactions on Speech and Audio Processing*, volume 8 (pp. 519–532).

O'Connor, J. & Arnold, G. (1961). *Intonation of colloqiual English*. London: Longman. 2nd edition 1973.

Öhman, S. (1967). Word and sentence intonation: A quantitative model. *Speech Transmission Laboratory - Quarterly Progress and Status Report*, 2, 20–54.

Ostendorf, M. & Ross, K. (1997). A multi-level model for recognition of intonation labels. In Y. Sagisaki, N. Campbell, & N. Higuchi (Eds.), *Computing Prosody: Computational Models for Processing Spontaneous Speech* chapter 19, (pp. 291–308). New York: Springer.

Palmer, H. (1922). *English intonation with systematic exercises*. Cambridge: Cambridge University Press.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT.

Pierrehumbert, J. & Hirschberg, J. (1990). *The meaning of intonational contours in the interpretation of discourse*, chapter 14, (pp. 271–311). Intentions in Communication. MIT Press: Cambridge, MA.

Pierrehumbert, J. B. (1983). Automatic recognition of intonation patterns. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics* (pp. 85–90). Cambridge, Massachusetts: MIT.

Pitrelli, J., Beckman, M., & Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP)*, volume 2 (pp. 123–126). Yokohama, Japan.

Plank, F. & Filimonova, E. (2000). The universals archive: A brief introduction for prospective users. *Sprachtypologie und Universalienforschung (STUF)*, 53, 109–123.

Rapp, S. (1996). Goethe for prosody. *Internet paper. http://www.ims.uni-stuttgart.-de/∼rapp/A609.ps.gz.*

Reetz, H. (1996). *Pitch Perception in Speech: A Time Domain Approach*. Dordrecht: Foris.

Reetz, H. (1998). Automatic speech recognition with features. Universität des Saarlands. Habilitationsschrift.

Reetz, H. (1999). *Artikulatorische und akustische Phonetik*. Trier: Wissenschaftlicher Verlag Trier.

Reyelt, M. & Batliner, A. (1994). *Ein Inventar prosodischer Etiketten für Verbmobil*. Technical report, Verbmobil Memo 33.

Reyelt, M., Grice, M., Benzmüller, R., Mayer, J., & Batliner, A. (1996). Prosodische Etikettierung des Deutschen mit ToBI. In D. Gibbon (Ed.), *Natural Language and Speech Technology, Results of the third KONVENS conference* (pp. 144–155). Berlin: Mouton de Gruyter.

Rossi, M. (2000). *Intonation: Past, Present, Future*, volume 15 of *Text, speech, and language technology*, chapter 2, (pp. 13–52). Kluwer Academic Publishers: Washington, DC.

Silverman, K. (1987). *The Structure and Processing of Fundamental Frequency Contours*. PhD thesis, University of Cambridge, Dordrecht.

Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P., & Hirschberg, J. (1992). ToBI: A standard scheme for labeling prosody. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP)* (pp. 867–879). Banff, Canada.

Sluijter, A. M. C. (1995). *Phonetic correlates of stress and accent*. Dordrecht: Foris.

Sluijter, A. M. C. & Van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100, 2471–2485.

Syrdal, A. K. & McGory, J. (2000). Inter-transcriber reliability of ToBI prosodic labeling. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)* (pp. 235–238). Bejing, China.

't Hart, J. & Cohen, A. (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics*, 1, 309–327.

't Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge Studies in Speech Science and Communication. Cambridge: Cambridge University Press.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn & K. K. Paliwal (Eds.), *Speech Coding and Synthesis* (pp. 495–518). Amsterdam: Elsevier Science.

Talkin, D. & Lin, D. (1997). Manual page from get_f0. In *ESPS Programs A-L* (pp. 1–4). Washington, DC: Entropic Research Laboratory, Inc.

Taylor, P. A. (1994). *A Phonetic Model of Intonation in English*. PhD thesis, University of Edinburgh, Bloomington, Indiana. Distributed by Indiana Linguistics University Club Publications.

ten Bosch, L. (1993). On the automatic classification of pitch movements. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 2 (pp. 781–784). Berlin, Germany.

Trager, G. & Smith, H. L. (1951). *An Outline of English Structure*. Norman, Oklahoma: Battenburg Press.

Uhmann, S. (1987). *Fokussierung und Intonation*. Phd thesis, University of Konstanz.

Vaissière, J. (1988). The use of prosodic parameters in automatic speech recognition. In H. Niemann, M. Lang, & G. Sagerer (Eds.), *Recent Advances in Speech Understanding and Dialog Systems*, volume 46 of *NATO ASI Series F* (pp. 71–99). Berlin, Germany: Springer.

Venditti, J. J. (1995). *Japanese ToBI Labelling Guidelines*. Technical report, Ohio State University.

Vereecken, H., Martens, J.-P., Grover, C., Fackrell, J., & Van Coile, B. (1998). Automatic prosodic labeling of 6 languages. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)* (pp. Paper Nr. 45). Sydney, Australia.

Viswanathan, V. R. & Russel, W. H. (1984). *Subjective and objective evaluation of pitch extractors for LPC and harmonic deviations vocoders*. Technical report, Bolt Beranek and Newman Inc. (Final report No. 5726).

Wahlster, W. (2000). *Verbmobil: Foundations of Speech-to-Speech Translations*. New York: Springer.

Wahlster, W., Bub, T., & Waibel, A. (1997). Verbmobil: The combination of deep and shallow processing for spontaneous speech translation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1 (pp. 71–74). Munich, Germany.

Waibel, A. (1988). *Prosody and Speech Recognition*. Research Notes in Artificial Intelligence. San Mateo, CA and London: Morgan Kaufman Publishers and Pitman.

Wightman, C. W. & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2, 469–481.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 91, 1707–1717.

Wightman, C. W., Syrdal, A. K., Stemmer, G., Conkie, A., & Beutnagel, M. (2000). Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2 (pp. 71–74). Bejing, China.