Goethe for Prosody

Stefan Rapp

Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart
Azenbergstr. 12, 70174 Stuttgart, Germany, e-mail: Stefan.Rapp@ims.uni-stuttgart.de

ABSTRACT

In this paper, we describe the way in which a recording of Goethe's "Die Leiden des jungen Werther" published on a multimedia CD-ROM [7] was made accessible for prosody research. The recording is interesting for prosody research because of its prosodic richness as it displays a large variety of registers and speaking styles. Application areas are: development of prosody models for German TTS, unsupervised learning of pitch accent types, corpus search for research on prosody-semantics and prosody-syntax interaction, and the study of more global prosodic parameters (speaking rate, pitch range) defining registers or speaking style. The four hour recording was segmented into phonemes, syllables and words using HMM speech recognition techniques [5, 13] together with a large pronunciation lexicon [1]. A part of speech tagger [14] was applied to the corpus to yield time aligned POS tags. The German adaptation of the tone sequence model of intonation used in Stuttgart [11, 6] inspired the parametrization of fundamental frequency. An intermediate phonetic representation layer is described that uses the syllable alignment to parametrize the F_0 contour into a superposition of three functions: a hyperbolic tangent, a gaussian and a constant.

1. INTRODUCTION

The age of multimedia and connectivity blesses us with Giga bytes of all kinds of text and hundreds of hours of speech—a lucky situation for researchers investigating language and speech. Recordings of radio plays and literature readings are anything but new, but with the advent of CD-ROM technology, the new thing is the compilation of speech *and* machine readable text on one medium. These media are usually very inexpensive compared to the effort it takes to make reasonable recordings in the lab. In this paper, we describe the way in which such a recording (that of Goethe's "Die Leiden des jungen Werther") was made accessible for prosody research. The four hour recording of the classic piece perfomed by a single professional male actor is published on a multimedia CD-ROM [7] along with corresponding orthographic text. The recording gives the opportunity to establish a rather large single speaker acted speech corpus. It is interesting for prosody research because of its prosodic richness. It displays a large variety of registers and speaking styles. The quality of the 8-bit 22.05 kHz recording is sufficient for estimation of syllable length, fundamental frequency and intensity, the prosodic parameters of prominence and phrasing. The next section describes the segmentation of the recorded speech into words, syllables and phonemes. The word segmentation is used to get time aligned part of speech tags, described in section 3, and the syllable segmentation is used for the F_0 parametrization as described in section 4. The paper ends with an indication of possible application areas and plans for future work.

2. AUTOMATIC SEGMENTATION

The orthographic text and the speech as found on the CD-ROM is organized in files that correspond roughly to the pages of the printed version. This parallel structuring enables the application of forced alignment to find the start and end points of individual speech segments. Some effort had to be made to accommodate different 'page breaks' in speech and text: As the speaker finishes a sentence that is only partially on the bottom of a page, for alignment the text of each page was extended by the 5 last lines of the previous page and the 5 first lines of the next page. Then, from the top of the text to the middle of the page and from interpunctation to interpunctation, increasing text chunks were marked optional (i.e., text from top to the first interpunctation is marked, text from top to the second interpunctation is marked etc.) Likewise, decreasing text chunks from around the middle of the page to the bottom were marked optional as well. As the alignment [13] uses HMM speech recognition technique [5], the procedure is able to chop off superfluous text (if it is marked as optional) at the top and bottom, making the page breaks the same for speech and text. By comparing the concatenation of the found word sequences ('speech pages') with the concatenation of the text pages, it is possible to find probably all recognition errors. (In the case of [7], 5 automatically detected small text chunks had to be manually reassigned to pages). This procedure gives us the transliteration of the speech, and the HMMs can be retrained for the speaker. By comparing the initial (speaker independent) segmentation with the segmentation produced by the retrained (speaker adapted) HMMs, considerable changes were observed. However, we will abstain from quantifying the differences in this paper. Retraining should especially be considered when planning to use such a database for concatenative speech synthesis.

In order to segment sub word units, one has to convert the text into phonemes. Grapheme phoneme conversion of [13] is based on a large 360 000 word form lexicon [1]. Only 4.82% OOV words had to be converted by the crude rule based conversion that follows the

binary search lexicon lookup. The syllabification and word stress information that is (among other things) included in [1] is used to generate a syllable segmentation from the phoneme alignment taking into account ambisyllabic consonants. In total, the 4 hour recording was almost fully automatically segmented into 39 421 words and 6 538 pauses, 63 743 syllables (of which 2 820 were different) and 169 970 phonemes. See fig. 1 for a tiny extract of the corpus.

3. PART OF SPEECH TAGGING

An extended version of Schmid's part of speech tagger [14] was applied to the corpus to yield time aligned POS tags. As long as robust parsers that can handle unrestricted text are not easily available, POS tags are used as an approximation to render linguistic function. It is useful to include POS tags in a corpus for, e.g., pitch accent recognition studies, because linguistic function is supposed to influence perception of pitch accents. Furthermore, it is useful to have time aligned POS tags for speech synthesis applications. For example, if considering concatenative speech synthesis, one might want to take segments only out of words having the same POS tag as the target word, or even only from words that appear in the same linguistic context. This may be helpful in finding an adequate degree of phonetic reduction which is dependent on linguistic function [4].

4. F_0 PARAMETRIZATION

The intonation model used in Stuttgart [11] inspired the parametrization of fundamental frequency. The Stuttgart System is an attempt to integrate the phonological analysis of German intonation done by C. Féry [6] and the ToBI labelling conventions [2, 8]. The system was developed as a tool within our overall aim which is to create a prosodic module for Discourse Representation Theory (DRT) [9]. DRT is a model-theoretic approach to discourse semantics which describes the interpretation of discourses as a dynamic two-level process. Hence, since our inventory of symbols is primarily motivated by phonological analysis and since the domain of processing these symbols is DRT, the criterion for describing fundamental frequency contours which is emphasized in our system is that only those intonational events should be labelled which are distinctive in the sense that one can assign them a function in the domain of discourse interpretation. A further consequence is, that the system reflects phonological pitch-accent linking rules which introduce a set of 'allotonic' accents, not considered further in this paper. Beyond that a small set of default symbols enriches the standard ToBI notation. These default labels are filled in with phonetic content according to autosegmental spreading and alignment conventions. Again, the reader is invited to take a look at [11] for further discussion.

The basic idea behind the parametrization is to find a phonetic description of the intonation contour that (a.) is computationally tractable and well defined and (b.) fits well to our phonological model of intonation in order to ease recognition and generation of pitch accents and boundary tones (see [15] for discussion of the advantages of a phonetic model of intonation as an intermediate representation). As the Stuttgart intonation model is phonology oriented, the basic timing structure for the phonetic layer, the F_0 parametrization, is the syllable segmentation found by the automatic alignment. The F_0 parametrization should ideally yield similar parametrizations for equal pitch accents or boundary tones and distinguishable parametrizations for differing pitch accents or boundary tones. In order to understand the chosen parametrization approach, we have to recall some basics of our labelling system. There are five standard pitch accents in our system: a L*H (rise), a H*L (fall), a L*HL (rise-fall/"latepeak"), a HH*L ("early peak") and a H*M (stylized contour). The first two accents, L*H and H*L are the most common ones. The last, H*M, is very infrequent, because there are strong phonological restrictions on it's use. It is produced for instance, when somebody is calling a person's name (vocative). We found a distribution of approximately 59 / 36 / 3 / 2 / 0 percent respectively in the material yet labelled at our institute. Because the H (for high) and L (for low) tones are each associated with one syllable and the accented syllable is marked with a star, it is clear from the accents' naming that the accented syllable's and the postaccented syllable's F_0 is involved in pitch accent perception of every accent type.[1] For two accents, HH*L and L*HL, more distant F_0 information seems relevant. For HH*L, information from the preaccented syllable contributes to distinctivity from H*L. A similar situation could be inferred from L*HL's naming and the F_0 of the syllable after the postaccented. But, as experience with manual labelling has shown, the "late peak" accent L*HL is mostly realised on the accented and postaccented syllable alone. It seems that cases where F_0 information from behind the postaccented syllable helps at distinguishing a L*H from a L*HL are quite rare. A more critical case is that of HH*L. Here, F_0 information from the preaccented syllable seems to play a role for disambiguating it from H*L, as manual labelling experience shows.

What follows for the parametrization is to consider a two syllable window, associated to the first (potentially accented) syllable. If the distinctivity of the parametrization between H*L and HH*L is not sufficient, a postprocessing stage can be applied taking into account the parametrization associated to the *pre*accented syllable (and only for these H*L vs. HH*L cases). In order to have convenient formula and easy to interpret parameters, the points in time for which F_0 measurement exist are linearly scaled to (-1,0) for the associated (first) syllable and to (0,1) for the following syllable. In addition, this scaling abstracts from syllable length.

It is clear that the parametrization should separate well between the two most frequent pitch accent types, H*L and L*H. When thinking of a nice, i.e., continuous and differentiateable, mathematical function that could express the tonal difference between two syllables, an inverse tangent and an hyperbolic tangent comes into mind. We chose an hyberbolic tangent as our first parametrization function. Three parameters α, β and γ control the shape of that function, allowing it to change its amplitude, to let it move a bit to the left or right and to adjust the steepness of the slope. Ideally, these three parameters should differentiate the four or five basic accents well, but experiments have shown that often there is a kind of peak (rise-fall) in the two syllable window where it is not clear if the $\tanh$ should

[1] Disregarding complete linking [6, 11] and a possible spreading of the starred H tone of the very rare H*M when the accented syllable is in antepenult position.

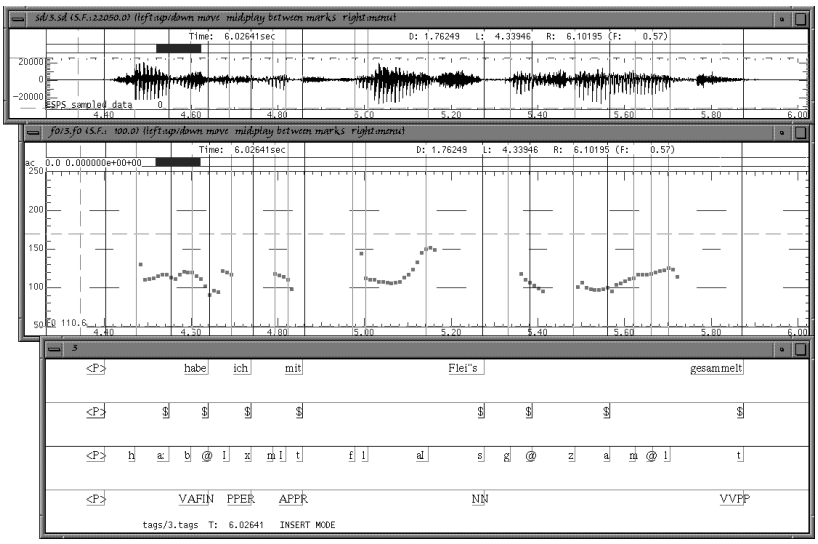


Figure 1: Results of automatic segmentation and POS tagging. Top to bottom: speech signal, F_0 contour, word- syllable- phoneme- and POS-tag labelling

model the rise or the fall. This happens when there are two bitonal accents on adjacent syllables, or in the case of a "late peak" L*HL accent realised on two syllables. A mathematical function that can describe a peak is e^{-x^2} which is well known from the probability density function of a normal distribution, a very natural choice, though. This is our second parametrization function. Again it is controlled by three parameters, δ, ε and ζ, describing the height of the peak, its temporal alignment (corresponding to the mean of a normal distribution) and its steepness (corresponding to the variance). Note that the temporal alignment of a peak is claimed to be relevant for pitch accent interpretation in German [10], a fact our intonation model tries to account for with the notion of the HH*L ("early peak"), H*L ("mid peak") and L*HL ("late peak") accents. As the two chosen parametrization functions don't add up to a speakers F_0 range, a third parametrization function is nescessary to lift the functions up into her or his range. Obvious choice is a straight line $ax + b$. Assuming declination of the overall contour of intonation, one could enforce a being negative and constant for all syllables of a phrase or even globally fixed. Since we (a.) do not want to decide beforehand how long phrases are, (b.) do not want interference with the first parametrization function and (c.) want to keep things simple, we have chosen a constant function as our third parametrization function. This constant, η, describes approximately the overall F_0 level in the two syllable window. In total, this is the phonetic parametrization function we exploit:

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

The seven parameters have the following phonetic interpretation:

- α correlates with the tonal difference of the accented and postaccented syllable from η. 2α is the height of the rise or fall.
- β correlates with the steepness of the rise or fall.
- γ correlates with the temporal alignment of the rise or fall.
- δ correlates with the height of the peak.
- ε correlates with the steepness of the peak.
- ζ correlates with the temporal alignment of the peak.
- η correlates with the overall F_0 level.

These phonetic parameters are obtained fully automatically by using MATLAB's `fmins` function which implements the Nelder-Meade simplex search [12, 3]. The idea behind this iterative optimization algorithm is to stepwise distort the parameters a bit such that the mean sqare error of F_0 values with respect to the phonetic parametrization function is minimized. In every step, the parameters are forced to be in a realistic range, e.g., η falls inside the speakers and the two syllables F_0 range. As the algorithm optimizes locally, it is not guaranteed that it finds the global optimum, i.e., best phonetic parametrization. To avoid misparametrizations, an initial guess is computed by simple heuristics that should hopefully lie in the vicinity of the global optimum.

A psycholinguistically more motivated approach might be to start the search with the parametrization of a prototypical accent of the kind one expects for the syllable in question solely from linguistic function of the preceding words. As we do not know yet what accent and boundary tone types to expect, we leave this issue to further research.

5. APPLICATION AREAS AND FUTURE WORK

The corpus is useful for general phonetic and phonological inquieries, e.g., to search for syllable, syllable onset, syllable nucleus, syllable rhyme or phoneme lengths in different contexts for one particular speaker. As the amount of data is rather large, it is presumably possible to train machine learning algorithms to predict these

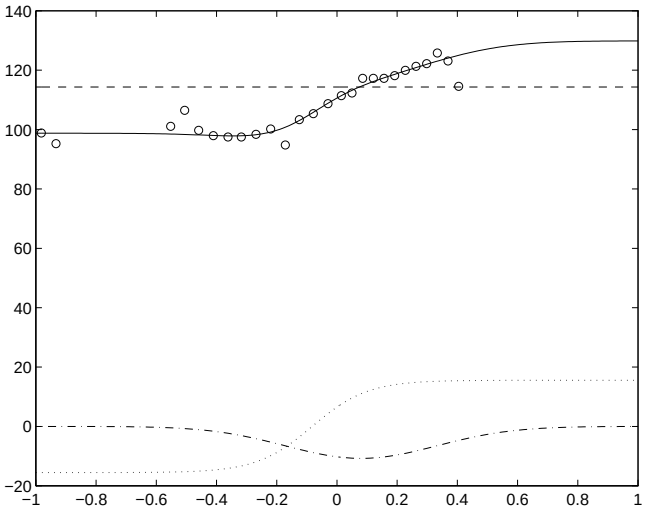


Figure 2: F_0 meassurement (circles) and parametrization function (solid line), consisting of: $\tanh$ (dotted line), e^{-x^2} (dashdot line), constant (dashed line) for the last two syllables of fig. 1.

lengths from given linguistic context (tags) and phonological context (word stress, syllable structure, adjacent phonemes) to yield a duration model for German text to speech. Another application area of the corpus is to use it as a database for corpus based concatenative speech synthesis. Studies that are currently projected at the IMS will show whether additional information about register or speaking style will be necessary to add. Investigation of global prosodic properties seems interesting. Unsupervised learning should be attempted to test clustering of text chunks by considering speaking rate, F_0 range and F_0 distribution from the corpus. These clusters might well correspond to registers and/or speaking style. Together with word alignment it is possible to investigate how discourse particles effect discourse structure and the mentioned global prosodic correlates.

Focusing on more local aspects of intonation, the corpus can serve for automatic extraction of a simple accent model by predicting the phonetic parametrization directly from tags. Such an approach could be used as a fallback strategy or a reference system to test more elaborated strategies of prosody generation exploiting, e.g., givenness and information structure.

The author's current main interest lies in pitch accent and boundary tone recognition, i.e., finding a mapping from the phonetic to the phonological layer. Preliminary experiments with a standard symbolic supervised machine learning program give encouraging results. Experimental material for this task consisted of 15 minutes of manually annotated radio news stories. The rules produced for pitch accent classification seem sound at first sight and justify to some extent the parametrization proposed in this paper. It is not clear yet to what extent these observations carry over to unsupervised learning techniques, sketched in the following: The F_0 parametrization spans a seven dimensional space. Specific pitch accents are supposed to appear in subspaces only. Again, clustering might help finding these regions. Finally a mixed strategy could be helpful: A larger corpus such as the corpus presented in this paper, might be used to find cases (or case clusters) which are represented insufficiently in the small, manually labelled news stories corpus—in analogy to part of speech tagging approaches [14], where unannotated large corpora are used to improve taggers trained on smaller ones.

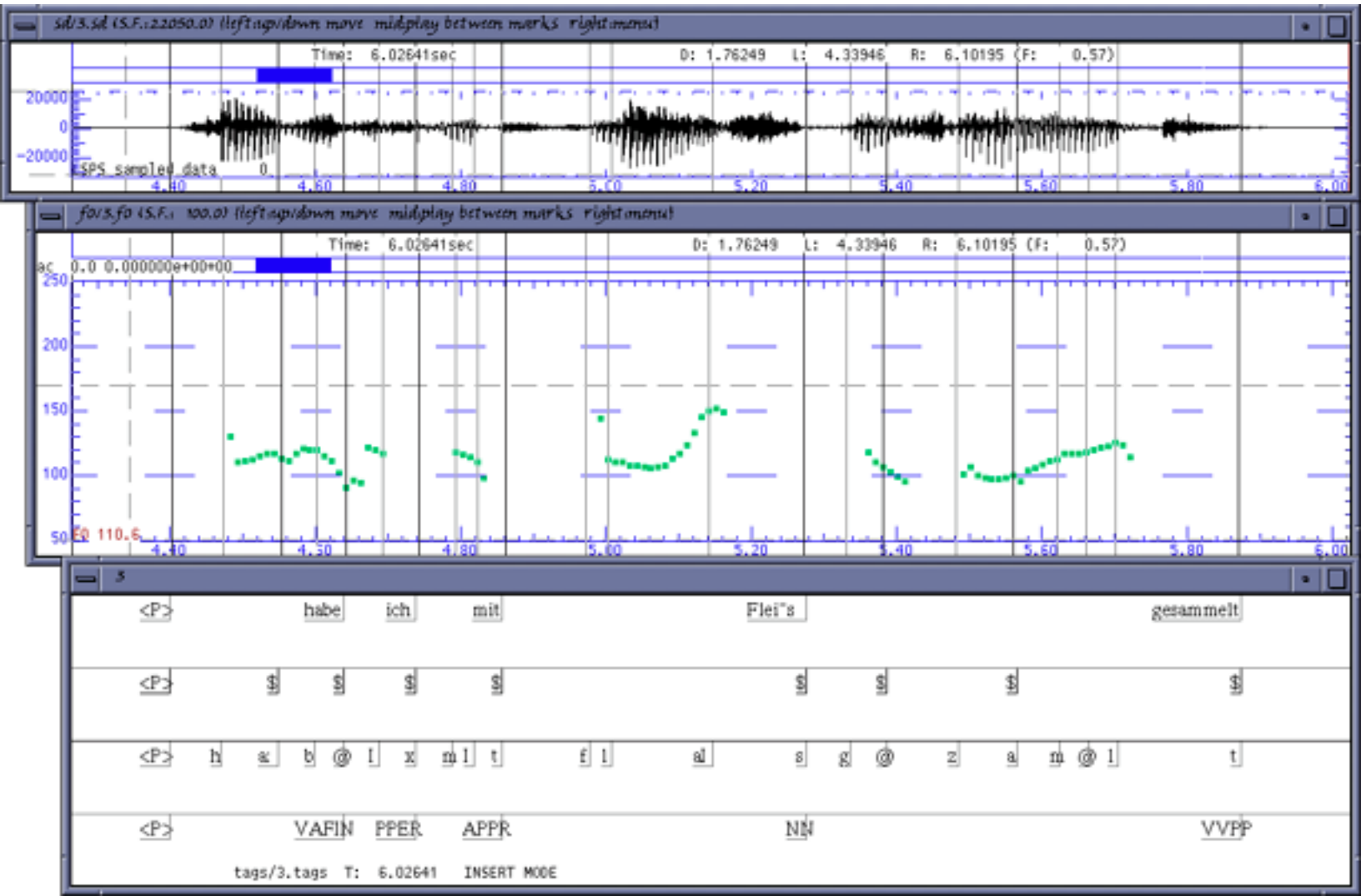
Goethe's "Werther" is one out of five CD-ROMs of classic literature recordings together with text that are currently available. The publishers plan to publish five pieces per season. As all extraction stages work almost automatically, the size of the corpus is easily extensible to tens or hundreds of hours of speech in the near future.

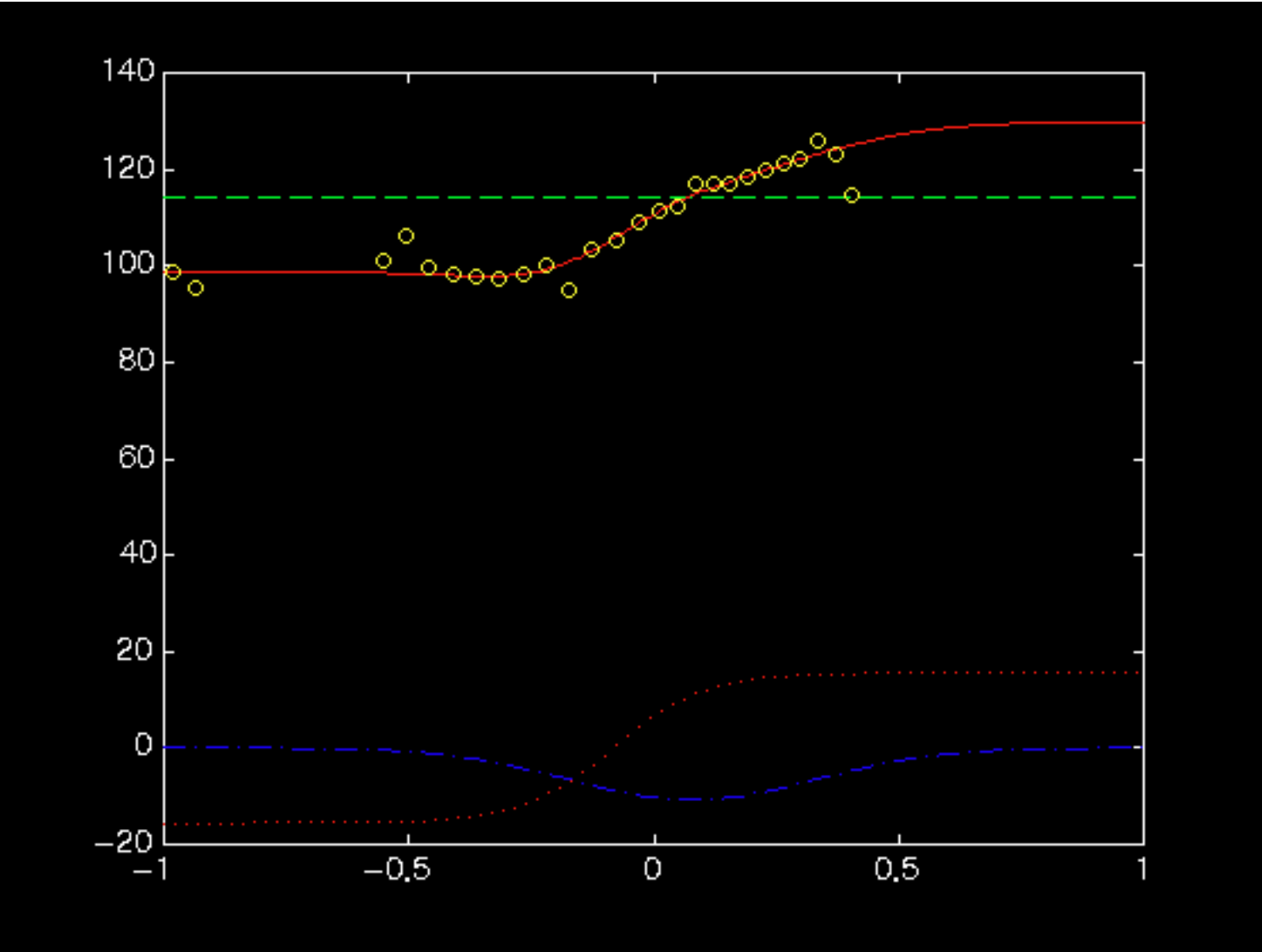
6. ACKNOWLEDGEMENTS

This work was supported by Deutsche Forschungsgesellschaft (DFG) within SFB 340 "Sprachtheoretische Grundlagen für die Computerlinguistik", project C4. I would like to thank Grzegorz Dogil, Jörg Mayer, Peter Regel-Brietzmann and Wolfgang Wokurek for fruitful discussions.

7. REFERENCES

1. R. H. Baayen, R. Piepenbrock, and H. van Rijn. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1993.
2. M. E. Beckman and G. M. Ayers. *Guidelines for ToBI labelling*, version 2.0 edition, February 1994.
3. J. E. Dennis, Jr. and D. J. Woods. New computing environments: Microcomputers in large-scale computing. pages 116–122. SIAM, 1987.
4. G. Dogil. Grammatical prerequisites to the analysis of speech style: Fast/casual speech. In D. Gibbon and H. Richter, editors, *Intonation, Accent and Rhythm: Studies in Discourse Phonology*, pages 91–119. de Gruyter, Berlin, 1984.
5. Entropic Research Laboratory, Inc, 600 Pennsylvania Avenue, Washington DC 20003. *HTK – Hidden Markov Model Toolkit*.
6. C. Féry. *German Intonational Patterns*. Niemeyer, Tübingen, 1993.
7. J. W. Goethe. *Die Leiden des jungen Werther*. Philipp Reclam jun., Stuttgart and Silver Spring, Berlin, Klassiker auf CD-ROM edition, 1995.
8. J. Hirschberg and M. E. Beckman. *The ToBI annotation conventions*, 1994.
9. H. Kamp and U. Reyle. *From discourse to logic*. Kluwer Academic Publishers, Dordrecht, 1993.
10. K. J. Kohler. A model of german intonation. Arbeitsberichte (AIPUK) 25, Institut für Phonetik und digitale Sprachverarbeitung Universität Kiel, 1991.
11. J. Mayer. Transcribing German intonation - the Stuttgart system. Manuscript, Universität Stuttgart, 1995. http://www.ims.uni–stuttgart.de/phonetik/joerg/labman/STGTsystem.html.
12. J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313.
13. S. Rapp. Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models / An aligner for German. In *Workshop "Integration of Language and Speech in Academia and Industry"*, Moscow, November 1995. ELSNET goes east and IMACS. http://www.ims.uni-stuttgart.de/~rapp/aligner.ps.gz.
14. H. Schmid. Improvements in part-of-speech tagging with an application to german. In *Proceedings of EACL SIGDAT-Workshop*, Dublin, Ireland, 1995. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.ps.gz.
15. P. A. Taylor. *A Phonetic Model of Intonation in English*. Indiana University Linguistics Club Publications, Bloomington, Indiana, 1994.


sd/3.sd (S.F.:22050.0) (left:up/down move midplay between marks right:menu)
Time: 6.02641sec
D: 1.76249 L: 4.33946 R: 6.10195 (F: 0.57)
20000
0
-20000
ESPS sampled data 0
4.40
4.60
4.80
5.00
5.20
5.40
5.60
5.80
6.00
fo/3.fo (S.F.: 100.0) (left:up/down move midplay between marks right:menu)
Time: 6.02641sec
D: 1.76249 L: 4.33946 R: 6.10195 (F: 0.57)
ac 0.0 0.000000e+00+00
250
200
150
100
50
F0 110.6
4.40
4.50
4.80
5.00
5.20
5.40
5.60
5.80
6.00
s
<P> habe ich mit Flei"s gesammelt
<P> $ $ $ $ $ $ $ $
<P> h a: b @ l x m l t f l al s g @ z a m @ l t
<P> VAFIN PPER APPR NN VVPP
tags/3.tags T: 6.02641 INSERT MODE


140
120
100
80
60
40
20
0
−20
−1
−0.5
0
0.5
1