A QUANTITATIVE DESCRIPTION OF GERMAN PROSODY OFFERING SYMBOLIC LABELS AS A BY-PRODUCT

Hansjörg Mixdorff and Hiroya Fujisaki

(mixdorff@tfh-berlin.de, fujisaki@te.noda.sut.ac.jp)

Dresden University of Technology and Science University of Tokyo

ABSTRACT

The prosodic quality of a text-to-speech system is important for the intellegibility and perceived naturalness of synthetic speech. In earlier works the author developed a linguistically motivated model of German intonation based on the quantitative Fujisaki model of the production process of F0. The current paper compares results yielded by automatic Fujisaki modeling with a GToBI-style anotation. On the accent level, a good correlation between tone labels and accent commands can be observed. On the phrase level, most level 3 and 4 break index boundaries are aligned with phrase commands whereas lower level boundaries are presumably marked with durational cues. Subsequently a regression model of syllable duration is introduced which permits to decompose the measured duration contour into an extrinsic and an intrinsic component.

1. INTRODUCTION

It is an undisputed fact that the intellegibility and perceived naturalness of synthetic speech strongly depends on the prosodic quality of a TTS system. Although some recent systems evade this problem by concatenating larger chunks of speech from a data base (see, for instance, [1]) which preserves the natural prosodic structure at least throughout the chunks chosen, the question of optimal unit-selection still calls for the development of prosodic models. Besides, the production process of prosody and the interrelation between the prosodic features of speech is far from being fully understood.

Earlier work by the authors was dedicated to a model of German intonation which uses the quantitative Fujisaki-model of the production process of F0 [2] for parametrizing F0 contours. The contour is described as a sequence of linguistically motivated tone switches, major rises and falls, which are modeled by onsets and offsets of accent commands connected to accented syllables or boundary tones. Prosodic phrases correspond to the portion of the F0 contour between consecutive phrase commands [3]. The model was integrated into the TU Dresden TTS system DRESS and proved to produce a high naturalness compared with other approaches [4]. Perception experiments, however, indicated flaws in the duration component of the synthesis system and raised the question how intonation and duration model should interact in order to achieve the highest prosodic naturalness possible. Most conventional systems like DRESS use separate modules for generating F0 and duration contours, modules which are often developed independently and use features derived from different data sources and environments. This ignores the fact that the natural speech signal is coherent in the sense that intonation and speech rhythm are co-occurrent and hence strongly correlated. As part of his post-doc thesis the first author of this paper decided to develop a prosodic module which is designed taking into account the coherence between melodic and rhythmic properties of speech. The model is henceforth to be called an 'integrated prosodic model'. For its F0 part this integrated prosodic model is still based on the Fujisaki model which is to be combined with a duration component.

2. SPEECH MATERIAL AND METHOD OF ANALYSIS

In the first phase of the project, a larger speech data base was analyzed in order to determine the statistically relevant input features of the integrated prosodic model. The corpus is part of a German corpus compiled by the Institute of Natural Language Processing, University of Stuttgart and consists of 48 minutes of news stories read by a male speaker [5]. The decision to use this database was made for several reasons: The data is real-life material and covers unrestricted informative texts produced by a professional speaker in a neutral manner. This speech material appears to be a good basis for deriving prosodic features for a TTS system which in most applications functions as a reading machine. As some of the news stories were recorded several times on the same day, intra-speaker consistency can be readily examined on the same data.

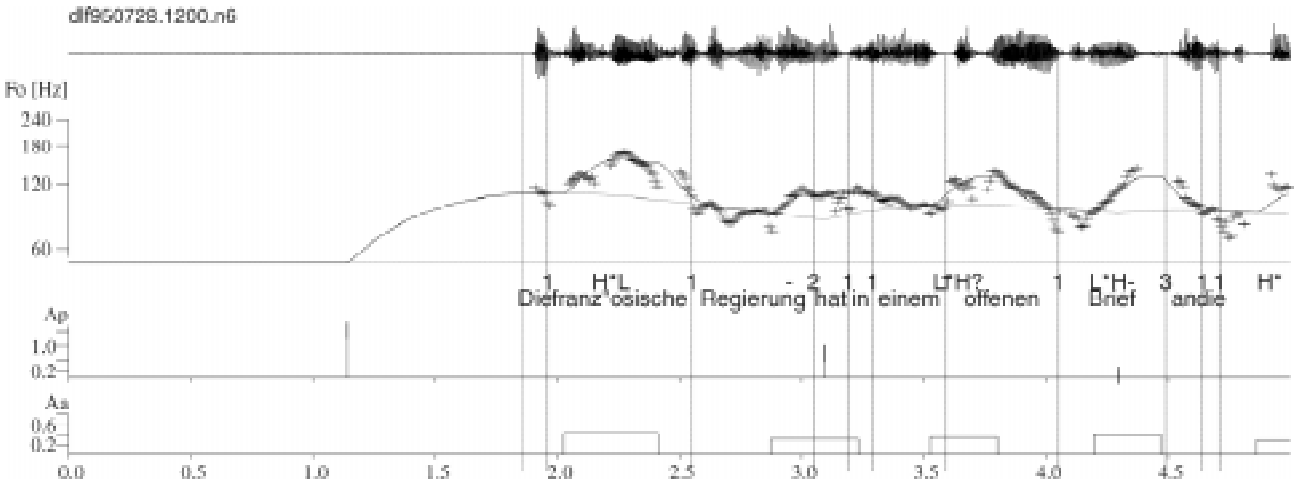

Figure 1: Example from the database. The figure displays from top to bottom: (1) the speech waveform, (2) the extracted (+-signs) and estimated (solid line) F0 contours, the ToBI labels and text of utterance, the underlying phrase commands (impulses) and accent commands (steps). In the utterance "Die französische Regierung hat in einem offenen Brief..." *"In an open letter, the French government..."* the second accent command marks a minor accent on 'Regierung'/*'government'* which was not assigned a ToBI-label.

The corpus contains boundary labels on the phone, syllable and word levels and linguistic annotations such as part-of-speech. Furthermore it is supplied with GToBI-labels following the Stuttgart System [6]. The Fujisaki-parameters were extracted applying a novel automatic multi-stage approach [7].

The present paper mainly concerns a side-aspect of the research project described, namely the comparison of automatically extracted Fujisaki parameters and ToBI labels as assigned by a human labeller. ToBI labels are symbolic in the sense that the labeller aims to describe the continuous F0 contour by a sequence of discrete tone labels, H for high tones, L for low ones. Syllables perceived as accented are assigned pitch accent labels ('starred' tones, such as L* and H*), prosodic boundaries are linked to boundary tone labels (such as H% or L%, for instance). Furthermore, the strength of prosodic boundaries is coded using break indices between 0 (clytic) and 4 (intonation phrase boundaries typically accompanied by pauses and boundary tones).

3. RESULTS

Figure 1 displays an example of analysis, showing from top to bottom: the speech waveform, the extracted and model-generated F0 contours, the ToBI tier, the text of the utterance, and the underlying phrase and accent commands.

3.1 Accentuation

The corpus contains a total number of 13151 syllables. Of the 2498 syllables labeled as accented ('H*L','L*H', etc.) 96.1% were found to be linked to accent commands, as well as 78% of the 859 syllables assigned boundary tone labels ('H%','L%'). 177 syllables marked with H% boundary tones receiving a separate accent command which is not linked to a preceding accent.

'Non-downstepped' accents exhibit a mean accent command amplitude of 0.28 against 0.21 for accents labeled as down-stepped.

The standard accent types 'H*L','L*H','HH*L' and 'L*HL' which account for 84% of the accent labels can be reliably identified by the alignment of the accent command with respect to the accented syllable, expressed as $T1_{rel}=(T1-t_{on})/dur$; and $T2_{rel}=(T2-t_{on})/dur$ where T1 denotes the accent command onset time, T2 the accent command offset time; t_{on} the syllable onset time and *dur* the accented syllable's duration. As can be seen from

Figure 2, for type 'H*L', mean $T1_{rel}$ and $T2_{rel}$ are -42% and 85%, and for type 'L*H' 50% and 170%, for instance. In a similar manner, the HH*L ('early high peak') and L*HL accent types (rise-fall / "late peak"), can be associated with the timing of the underlying accent command.

A considerable number of accented syllables (N=444) was detected which had not been assigned any accent labels by the human labeller. Figure 1 shows such an instance where in the utterance "Die Fran'zösische Re'gierung hat in einem 'offenen 'Brief..."-*"In an 'open 'letter, the 'French 'government..."* , an accent command was assigned to the word 'Re'gierung', but not a tone label. Other cases of unlabeled accents were incidently accented word accent syllables in by default unaccentable functions words.

3.2 Phrasing

About 54.8% of break index 3- and 96.2% of break index 4-labeled-boundaries are aligned with the onset of a phrase command, with a mean phrase command magnitued *Ap* of 0.67 and 1.32, respectively.

In order to separate intra-sentence from inter-sentence boundaries more consistently, a distinction not expressed by the BI3 and 4 labels, boundaries were post-labeled with default punctuation marks, i.e. periods and commas. Subsequently we found that all inter-sentence-boundaries are aligned with the onset of a phrase command. 68% of all intra-sentence

boundaries exhibit a phrase command, with the figure rising to 71% for 'comma-boundaries'. The mean phrase command magnitude for intra-sentence boundaries, inter-sentence-boundaries and paragraph onsets amounts to 0.8, 1.68 , and 2.28 respectively, which shows that *Ap* is a good indicator for boundary strength.

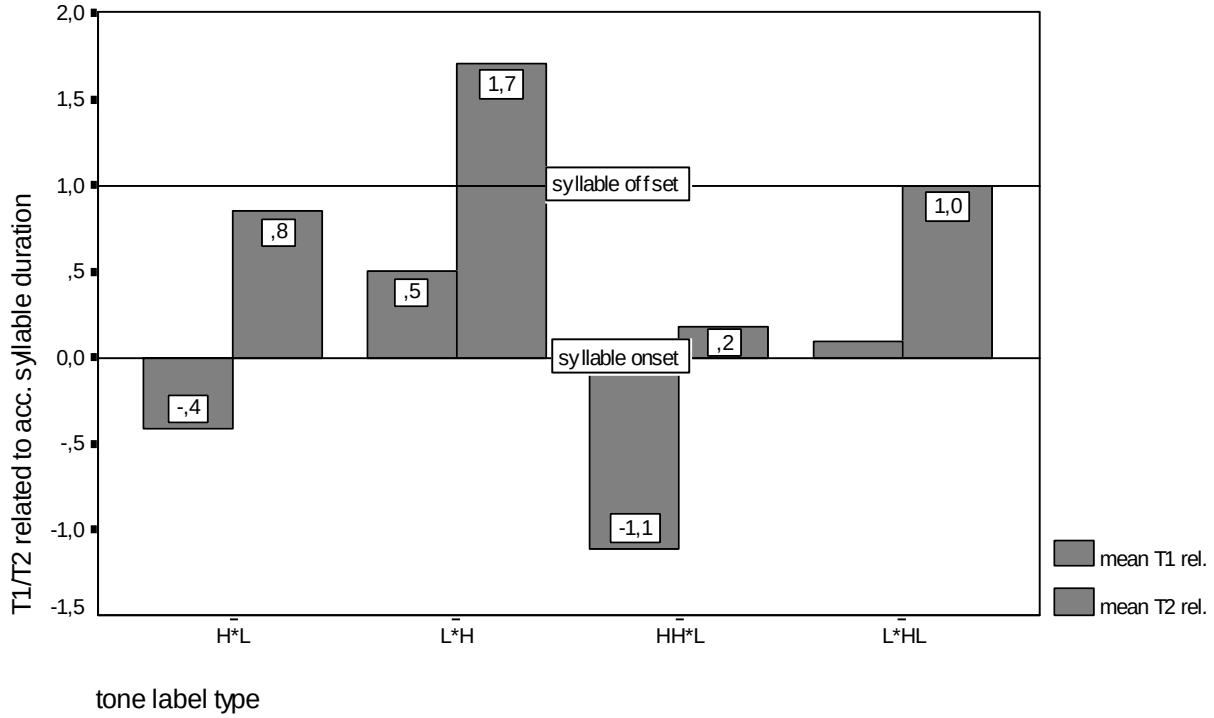

Figure 2: Timing characteristics of accent commands assigned to various ToBI accent label types expressed by T1 and T2 in relation to the accent syllable duration. H*L type accent commands start early before the syllable onsets, while L*H type accent commands start in the middle of the accented syllable.

About 80% of prosodic phrases contain 13 syllables or less. Hence phrases in the news utterances examined are considerably longer than the corresponding figure of 8 syllables found in [3] for simple readings. This effect may be explained by the higher complexity of the underlying texts, but also by the better performance of the professional announcer.

4. A PRELIMINARY MODEL OF SYLLABLE DURATION

As we saw in the preceding section, the labeling accuracy of the automatic procedure is quite high on the accent level and can be successfully used to determine tone labels without the loss of quantitative information incurred by a purely symbolic representation. The detection of lower level phrase boundaries, however, obviously requires the evaluation of additional features such as pausing and pre-boundary lengthening. A regression model of the syllable duration was hence developed which permits to decompose the duration contour into an 'intrinsic' part related to the syllable structure and a second, 'extrinsic' part related to accentuation and boundary influences. The most important extrinsic factors were found to be (1) the degree of accentuation (with the categories 0: 'unstressed', 1: 'stressed, but unaccented', 2: 'accented', where 'accented' denotes a syllable that bears a tone switch) and (2) the strength of the prosodic boundary to the right of a syllable, accounting for a total 35% of the variation in syllable duration. Pre-boundary lengthening, for instance, is therefore reflected by local maxima of the 'extrinsic' contour. The number of phones - as could be expected - proves to be the most important intrisic factor, followed by the type of the nuclear vowel to be either the reduction-prone schwa or non-schwa. These two features alone account for 36% of the variation explained. Figure 3 displays an example of a smoothed syllable duration contour (solid line) decomposed into an intrinsic (dotted line) and extrinsic (dashed line) component.

Compared with other duration models, the model presented here still incurs a considerable prediction error as it yields a correlation of only 0.79 between observed and predicted syllable durations, against a value of 0.85 in [8], for instance. Possible reasons for this shortcoming include the following:

- the duration model is not hierarchical, as factors from several temporal domains (i.e. phonemic, syllabic and phrasal) are superimposed on the syllabic level, and the detailed phone structure is (not yet) taken into account
- syllabification and transcription information in the database is often erroneous, especially for foreign names and infrequent compound words which were not transcribed using a phonetic dictionary, but by applying default grapheme-to-phoneme rules.

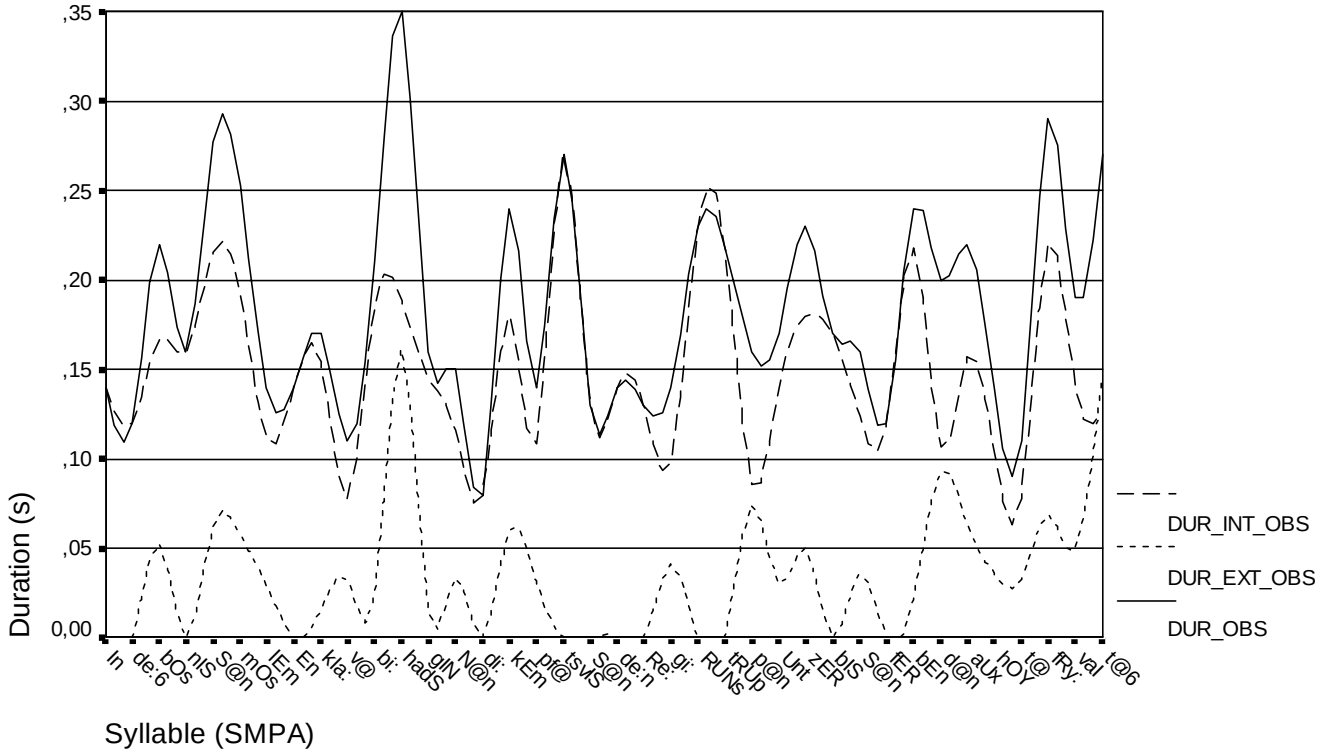

Figure 3: Example of smoothed syllable duration contours for the utterance "In der bosnischen Moslem-Enklave Bihac gingen die Kämpfe zwischen den Regierungstruppen und serbischen Verbänden auch heute früh weiter."-*"In the Bosnian Muslim-enclave Bihac, fights between the government troops and Serbian formations still continued this morning."* The solid line indicates measured syllable duration, the dashed line intrinsic syllable duration and the dotted line extrinsic syllable duration. At the bottom, the syllabic SMPA-transcription is displayed.

5. SUMMARY

The current paper compared a (symbolic) GToBI annotation with automatically extracted (quantitative) Fujisaki model parameters. Results show that on the accent level there is a strong correlation between ToBI accent labels and accent commands determined. As accent types can be readily identified by the relative onset and offset times of accent commands, the approach presented could be applied for performing a 'first guess' of ToBI-labels unbiased by the 'selectivity' of a human labeller.

Higher level boundaries are marked by the onset of phrase commands, whereas the detection of lower level boundaries will require the evaluation of durational factors. For this purpose a syllable duration model was introduced.

Besides the improvement of the syllable duration model, work is in progress for combining intonation and duration model into the integrated prosodic model.

6. REFERENCES

[1] Stöber K.; Portele T.; Wagner P.; Hess W. (1999): Synthesis by Word Concatenation. *Proceedings of EUROSPEECH '99*., vol. 2, pp. 619-622. Budapest 1999.

[2] Fujisaki, H. and Hirose, K. “Analysis of voice fundamental frequency contours for declarative sentences of japanese”. In *Journal of the Acoustical Society of Japan (E)*, 5(4): pp. 233-241, 1984.

[3] Mixdorff, H. Intonation Patterns of German - Model-based. Quantitative Analysis and Synthesis of F0-Contours. PdD thesis TU Dresden, 1998 (`http://www.tfh-berlin.de/~mixdorff/thesis.htm`).

[4] Mixdorff, H. and Mehnert, D. “Exploring the Naturalness of Several German High-Quality-Text-to-Speech Systems”, *Proceedings of Eurospeech '99*, vol.4, pp.1859-1862, Budapest, Hungary, 1999.

[5] Rapp, S. Automatisierte Erstellung von Korpora für die Prosodieforschung, PhD thesis Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. 1998.

[6] Mayer, J. Transcription of German Intonation: The Stuttgart System. Technischer Bericht, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. 1995.

[7] Mixdorff, H. “A novel approach to the fully automatic extraction of fujisaki model parameters”. In *Proceedings ICASSP 2000*, vol. 3, pp. 1281-1284, Istanbul, Turkey, 2000.

[8] Zellner-Keller, B.: “Prediction of Temporal Structure for Various Speech Rates”. In N. Campbell (ed.) Volume on Speech Synthesis. Springer-Verlag, 1998.