Symbolic versus quantitative descriptions of F0 contours in German: Quantitative modeling can provide both

Hansjörg Mixdorff and Hiroya Fujisaki

Dresden University of Technology
Mommsenstr. 13
01062 Dresden, Germany
mixdorff@tfh-berlin.de

Science University of Tokyo
2641 Yamazaki
Noda, 278 Japan
fujisaki@te.noda.sut.ac.jp

ABSTRACT

The problem of adequately describing F0 contours is far from being solved. Although symbolic representations such as the ToBI-system appear attractive and have been strongly promoted recently, they neither capture F0 contours in a way that permits their reproduction from the labels, a demand resulting from TTS, nor are the labels phonological in the strict sense. The current paper compares results yielded by automatic quantitative modeling with a GToBI-style annotation. 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. The results of the study indicate that Fujisaki model parameters, while preserving the F0 contour information, can as well be used for deriving a ToBI-style representation.

1. Introduction

In recent years symbolic representations of F0 contours such as ToBI [1] have become increasingly popular. These representations appear to provide a consistent labeling framework based on a limited set of rules which claim to be phonologically motivated. This claim which has been sometimes questioned shall not be further discussed in the scope of this paper which is mainly concerned with the intellegibility and perceived naturalness of synthetic speech. Predicting F0 contours obviously requires quantitative information about the F0 contour underlying an utterance and its connection to the segmental string which is not preserved in the symbolic ToBI labels.

In recent years the authors developed a model of German intonation which uses the quantitative Fujisaki model of the production process of F0 [2] for parametrizing F0 contours. This model was originally designed for the common Japanese. In the case of German, the F0 contour is described as a sequence of linguistically motivated tone switches [3], major rises and falls, which are modeled by onsets and offsets of accent commands connected to accented syllables or boundary tones. Accents are classified depending on their communicative function using the intoneme paradigm [4], prosodic phrases correspond to the portion of the F0 contour between consecutive phrase commands [5]. The model was integrated into the TU Dresden TTS system DRESS and proved to produce a high naturalness compared with other approaches [6]. Perception experiments, however, revealed 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 TTS systems like DRESS use separate modules for generating F0 and duration contours. These modules are often developed independently and are based on 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 is working on 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

For extracting prosodic parameters, 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 [7]. 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.

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 [8]. The Fujisaki-parameters were extracted applying a novel automatic multi-stage approach [9].

The present paper focusses on 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).

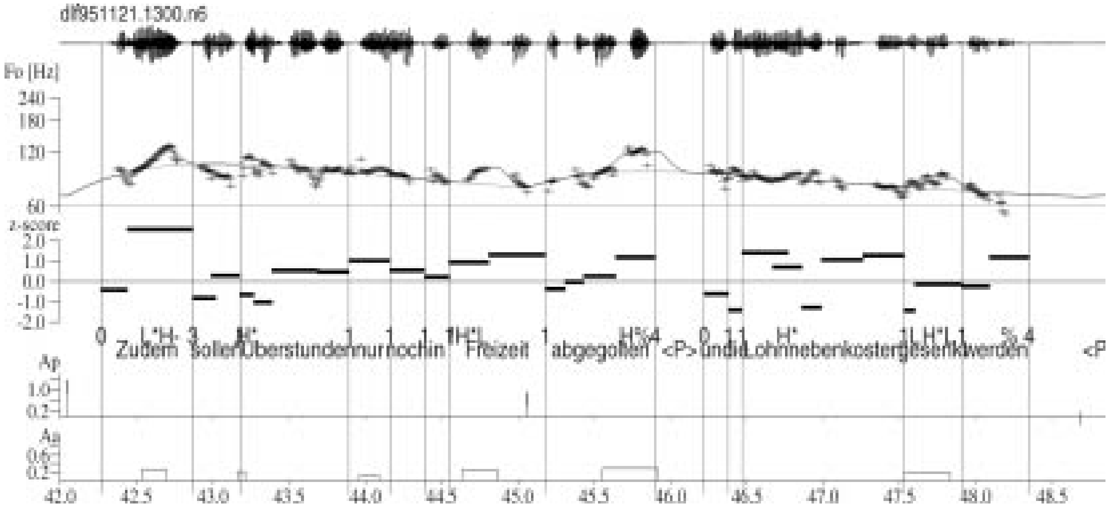

Figure 1: An example of analysis from the data base. The figure displays from top to bottom: (1) the speech waveform, (2) the extracted (+-signs) and estimated (solid line) F0 contours, the duration contour in terms of the syllable z-score, ToBI labels and text of utterance, the underlying phrase commands (impulses) and accent commands (steps). In the utterance "Zudem sollen Überstunden nur noch in Freizeit abgegolten und die Lohnnebenkosten gesenkt werden." ***"Furthermore, overtime will be compensated by time off in lieu only, and additional costs of wages are to be reduced."*** **the third accent command marks a minor accent on 'nur'/'only' which was not assigned a ToBI-label.**

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 duration contour in terms of the syllabic z-score drawn as horizontal lines of the length of the respective syllable, the ToBI tier, the text of the utterance, and the underlying phrase and accent commands.

3.1 Accent Assignment

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 were marked with H% boundary tones receiving a separate accent command which is not linked to a preceding accent (see, for instance, the accent command assigned to the word 'abgegolten' in Figure 1). Accents immediately preceding a boundary are found to be significantly stronger (with a mean accent command amplitude Aa of 0.38) than non-boundary accents with a mean Aa of 0.26.

'Non-downstepped' accents (98.0% of all accent labels) exhibit a mean accent command amplitude of 0.28 against 0.21 for accents labeled as downstepped. Furthermore, accents marked as uncertain ('?', 1.9 % of all accent labels) exhibit significantly lower Aa than those labeled with certainty (0.21 against 0.28). This indicates that it is the assessment of weaker accents that usually poses problems to the labeller.

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 $T1_{dist}$ =($T1-t_{on}$); and $T2_{dist}$=($T2-t_{off}$) where T1 denotes the accent command onset time, T2 the accent command offset time; t_{on} the syllable onset time and t_{off} the accented syllable's offset time. For type 'H*L', mean $T1_{dist}$ and $T2_{dist}$ are -60 ms and -37 ms, and for type 'L*H' 132 ms and 168 ms, respectively. In a similar manner, the HH*L ('early high peak') (-215 ms/-172 ms) and L*HL accent types (rise-fall / "late peak") (27 ms/-68 ms), can be associated with the timing of the underlying accent command.

A considerable number of syllables (N=444) exhibiting accent commands had not been assigned any accent labels by the human labeller. Figure 1 shows such an instance where in the utterance "Zudem sollen Überstunden nur noch in Freizeit abgegolten und die Lohnnebenkosten gesenkt werden." - *"Furthermore, overtime will be compensated by time off in lieu only, and additional costs of wages are to be reduced."*, an accent command was assigned to the word 'nur', but not a tone label. Closer analysis shows that labels are mainly missing when accents are relatively weak or in the case of secondary accents of longer compund words.

3.2 Phrase Boundaries

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

It must be stated, however, that the assignment of BIs by the labeller was sometimes inconsistent as boundaries with quite different prosodic cues and syntactic depths were assigned the same BI. Prosodic cues observed for boundaries include declination line resets - as triggered by phrase commands -, pauses, boundary tones and pre-boundary lengthening, the latter sometimes being the only cue at BI3 prosodic boundaries. As can be seen in Figure 1, the BI 3 boundary after 'Zudem' is mainly signaled by a durational cue (z-score=2.8 on the syllable 'dem'), whereas the BI4 boundaries after 'abgegolten' und 'werden' exhibit durational cues, as well as pauses. The sentence-medial boundary is also preceded by a phrase command adjusting the declination line and a high boundary tone connected to an accent command.

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.

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 [5] 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.

5. SUMMARY

The current paper compared a (symbolic) GToBI annotation with automatically extracted (quantitative) Fujisaki model parameters. The automatic procedure reliably assigns accent commands to the vast majority of the syllables labeled as accented and can be successfully used to determine tone labels without the loss of quantitative information incurred by a purely symbolic representation. As accent types can be readily identified by the onset and offset times of accent commands in relation to the accent syllable, the approach presented could be applied for performing a 'first guess' of ToBI-labels unbiased by the 'selectivity' of a human labeller. In the case of high boundary tones the problem arises of how to distinguish them from accents, as they often simply induce a lengthening of the accent command assigned to the last pre-boundary accent. Only in a minority of cases boundary tones were assigned a proper accent command.

Higher level boundaries are marked by the onset of phrase commands, whereas the detection of lower level boundaries will require the evaluation of additional features such as pausing and pre-boundary lengthening. Work is in progress towards the integration of these features into a general prosodic model of German.

REFERENCES

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Osterndorf, C.Wightman, P.Price, J. Pierrehumbert, and J. Hirschberg (1992): Tobi: A standard for labeling English prosody. In *Proceedings of ICSLP 1992*, Banff, 867-870.

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

[3] A.V. Isačenko and H.J. Schädlich (1964). *Untersuchungen über die deutsche Satzintonation*. Akademie-Verlag, Berlin.

[4] E. Stock and C. Zacharias (1982). *Deutsche Satzintonation*. VEB Verlag Enzyklopädie, Leipzig.

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

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

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

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

[9] Mixdorff, H. (2000): A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proceedings of ICASSP 2000*, vol. 3, 1281-1284, Istanbul, Turkey.