MFGI, a Linguistically Motivated Quantitative Model of German Prosody

Hansjörg Mixdorff *(mixdorff@tfh-berlin.de)*

Dresden University of Technology, D-01062 Dresden, Germany

Introduction

The intellegibility and perceived naturalness of synthetic speech strongly depends on the prosodic quality of a TTS system. Although some recent systems avoid this problem by concatenating larger chunks of speech from a database (see, for instance, Stöber et al., 1999), an approach which preserves the natural prosodic structure at least throughout the chunks chosen, the question of optimal unit-selection still calls for the development of improved prosodic models. Furthermore, the lack of prosodic naturalness of conventional TTS systems indicates that the production process of prosody and the interrelation between the prosodic features of speech is still far from being fully understood.

Earlier work by the author was dedicated to a model of German intonation which uses the well-known quantitative Fujisaki model of the production process of F0 (Fujisaki & Hirose, 1984) for parametrizing F0 contours, the **M**ixdorff-**F**ujisaki Model of **G**erman **I**ntonation (short **MFGI**). In the framework of MFGI, a given F0 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 so-called *boundary tones*. Prosodic phrases correspond to the portion of the F0 contour between consecutive phrase commands (Mixdorff, 1998). MFGI was integrated into the TU Dresden TTS system DRESS (Hirschfeld, 1996) and proved to produce a high naturalness compared with other approaches (Mixdorff & Mehnert, 1999).

Perception experiments, however, indicated flaws in the duration component of the synthesis system and gave rise to the question how intonation and duration model should interact in order to achieve the highest prosodic naturalness possible. Most conventional systems like DRESS employ separate modules for generating F0 and segment durations. These modules 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 author of this article 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 still relies on the Fujisaki model which is combined with a duration component.

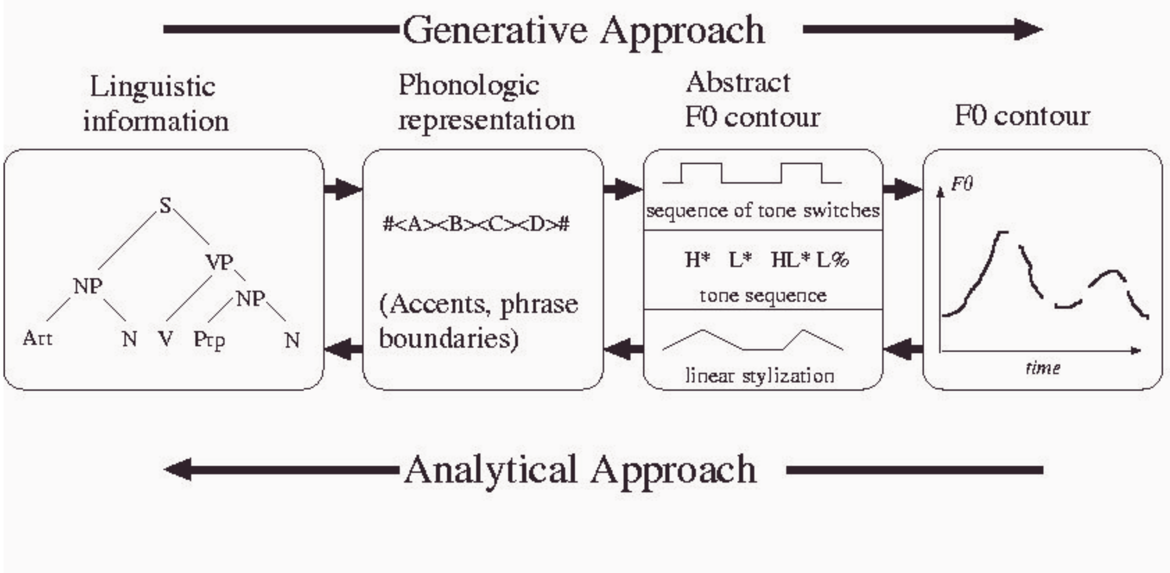

Figure 1. The role of intonation models as links between linguistic structures and their acoustic manifestations in the F0 contours.

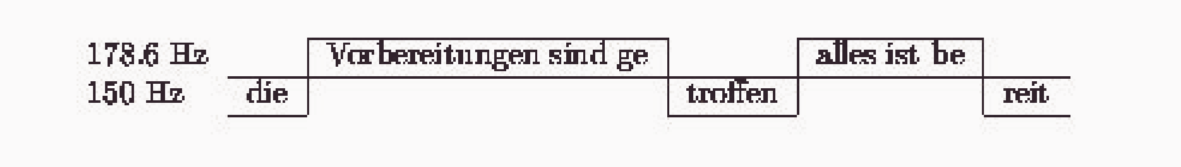

Figure 2. Illustration of the splicing technique used by Isačenko. Every stimulus is composed of chunks of speech monotonized either at 150 or 178.6 Hz.

Intonation Models

Figure 1 illustrates the role of intonation models as links between linguistic structures and their acoustic manifestation, the F0 contour. In principle, two general methods can be distinguished: One type of model (from the left to the right in the figure) deduces a phonological description from a linguistic structure (typically, the syntactic surface structure) specifying accent levels and phrase boundaries, transforming these into an abstract description of the F0 contour which is then by application of phonetic rules converted into the actual F0 contour. This type of model generates F0 contours from higher-level linguistic information and hence the method can be called "a generative approach" (see, for instance, Pierrehumbert, 1980, or Kohler, 1991). The opposite approach (from the right to the left) aims at abstracting from the the observable F0 contour by means of approximation techniques, yielding phonologically relevant basic elements. These are then used to infer linguistic units and structures. Since this approach is based on the analysis of the F0 contour observed, either mathematically, graphically or auditorily, it will be called an "analytical approach" (see, for instance, Adriaens; 1991, Hirst, 1993, Mertens & d'Alessandro, 1995). Hence an **ideal** intonation model by nature needs to be bi-directional, it

(1) possesses analytical power for extracting linguistic information from the F0 contour, and (2) generative power for synthesizing a contour from a set of linguistic input parameters.

The author believes that MFGI has the potential for being bi-directional. It makes use of a mathematical formulation for parametrizing a given F0 contour yielding a parsimonious set of parameters which can then be related to linguistic units and structures (analysis). On the other hand, using a set of linguistic units and structures, model parameters can be predicted and converted into a naturally sounding F0 contour (generation). Since the Fujisaki model proper is language independent, constraints must be defined as to its application to German. These constraints which differ from the implementation by Möbius et al. (1993), for instance, are based on earlier works on German intonation discussed in the following section.

Linguistic Background of MFGI

The early work by Isačenko (1964) is based on perception experiments using synthesized stimuli with extremely simplified F0 contours. These were designed to verify the hypothesis that the syntactic functions of German intonation can be modeled using tone switches between two constant F0 values connected to accented, so-called *ictic* syllables and *pitch interrupters* at syntactic boundaries.

The stimuli were created by monotonizing natural utterances at two constant frequencies and splicing the corresponding tapes at the locations of the tone switches (see Figure 1 for an example). The experiments showed a high consistency in the perception of intended syntactic functions in a large number of subjects.

The tutorial on German sentence intonation by Stock and Zacharias (1982) further develops the concept of tone switches introduced by Isačenko. The authors define phonologically distinctive elements of intonation called *intonemes*. Intonemes are characterized by the occurrence of a tone switch at an accented syllable. Depending on their communicative function, the following classes of intonemes are distinguished:

- *Information intoneme I↓* Declarative-final accents, falling tone switch. Conveying a message.
- *Contact intoneme C↑* Question-final accents, rising tone switch. Establishing contact.
- *Non-terminal intoneme N↑* Non-final accents, rising tone switch. Signaling non-finality.

Any intonation model for TTS requires information as to the appropriate accentuation and segmentation of an input text. In this respect, Stock and Zacharias' work is extremely informative as it provides default accentuation rules (word accent, phrase and sentence accents), and rules for the prosodic segmentation of sentences into *accent groups*.

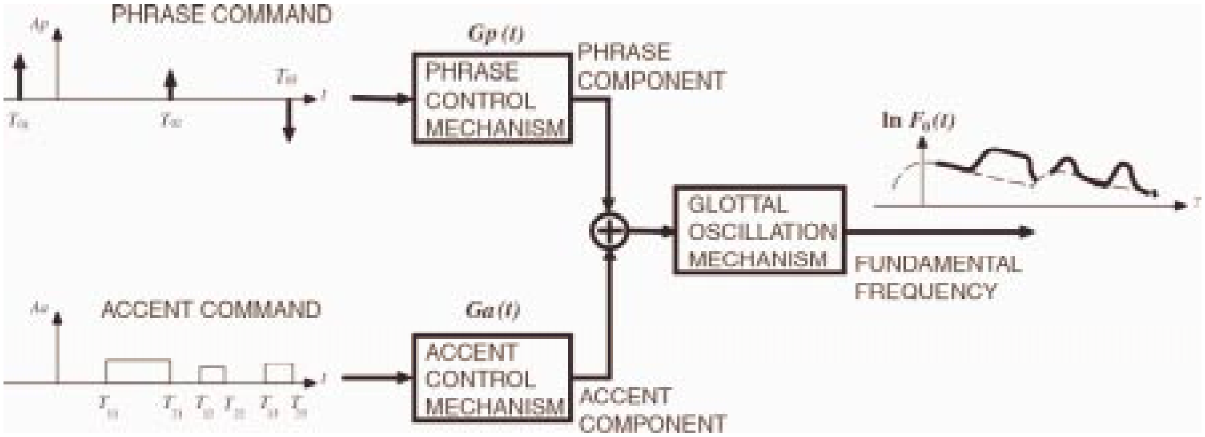

Figure 3. Block diagram of the Fujisaki model (Fujisaki and Hirose, 1984)

The Fujisaki Model

The mathematical formulation used in MFGI for parametrizing F0 contours is the well-known Fujisaki model. Figure 3 displays a block diagram of the model which has been shown to be capable of producing close approximations to a given contour from two kinds of input commands: phrase commands (impulses) and accent commands (stepwise functions). These are described by the following model parameters (henceforth referred to as *Fujisaki parameters*):

Ap: phrase command magnitude; *T0*: phrase command onset time; *α*: time constant of phrase command; *Aa*: accent command amplitude; *T1*: accent command onset time; *T2*: accent command offset time; *β*: time constant of accent command; *Fb*, the 'base frequency', denoting the speaker-dependent asymptotic value of F0 in the absence of accent commands.

The phrase component produced by the phrase commands accounts for the global shape of the F0 contour and corresponds to the declination line. The accent commands determine the local shape of the F0 contour, and are connected to accents. The main attraction of the Fujisaki model is the physiological interpretation which it offers for connecting F0 movements with the dynamics of the larynx (Fujisaki, 1988), a viewpoint not inherent in any other of the currently used intonation models which mainly aim at breaking down a given F0 contour into a sequence of 'shapes' (e.g. Taylor, 1995, Portele, 1995).

MFGI's Components

Following Isačenko and Stock, an F0 contour in German can be adequately described as a sequence of tone switches. These tone switches can be regarded as basic intonational elements. The term *intoneme* proposed by Stock shall be adopted to classify those elements that feature tone switches on accented syllables. Analogously with the term *phoneme* on the segmental level, the term *intoneme* describes intonational units that are quasi-discrete and denote phonological contrasts in a language. Although the domain of an intoneme may cover a larger portion of the F0 contour, its characteristic feature, the tone switch, can be seen as a discrete event. By means of the Fujisaki model intonemes can be described not only qualitatively, but quantitatively, namely by the timing and amplitude of accent commands to which they are connected. As presented in the preceding section, there are three classes of intonemes: The *information intoneme I↓*, the *contact intoneme C↑*, and the *non-terminal intoneme N↑* where the arrow indicates the direction of the tone switch. Since the I-intoneme may also occur in utterances of questions, it does not stand in a statement/question opposition with the C-intoneme.

Analysis of natural F0 contours (Mixdorff, 1998) indicated that further elements - not necessarily connected to accented syllables - are needed which occur at prosodic boundaries and will be called *boundary tones* (marked by *B↑*), a term proposed by Pierrehumbert (1980).

Further discussion is needed as to the question how the portions of the F0 contour pertaining to a particular intoneme can be delimited. In an acoustic approach, for instance, an intoneme could be defined as starting with its characteristic tone switch and extending until the characteristic tone switch of the following accented syllable. In the present approach, however, a division of the F0 contour into portions belonging to meaningful units, i.e., words or groups of words is favored, as the location of accented syllables is highly dependent on constituency, i.e. the choice of words in an utterance and the location of their respective word accent syllables. Unlike other languages German has a vast variety of possible word accent locations for words with the same number of syllables. Hence the delimitation of intonemes is strongly influenced by the lexical and syntactic properties of a particular utterance. We therefore follow the notion of *accent group* as defined by Stock, namely the grouping of clytics around an accented word as in the following example: "Ich s'ah ihn // mit dem F'ahrrad // über die Br'ücke fahren."-*"I saw him ride his bike across the bridge"*. where ' denotes accented syllables and // denotes accent group boundaries.

Analysis of natural F0 contours showed that any utterance is invariably preceded by a phrase command, and further commands in utterance-medial positions are usually linked with major prosodic boundaries. Hence, the term *prosodic phrase* denotes the part of an utterance between two consecutive phrase commands. It must be noted that since the phrase component possesses a finite time constant, a phrase command usually occurs shortly before the segmental onset of a prosodic phrase, typically a few hundred ms. The phrase component of the Fujisaki model is interpreted as a *declination component* from which rising tone switches depart and to which falling tone switches return.

Speech Material and Method of Analysis

In its first implementation, for generating Fujisaki parameters from text, MFGI relied on a set of rules (Mixdorff, 1998, p. 238 ff.). These were developed based on the analysis of a corpus which was not sufficiently large for employing statistical methods, such as neural networks or CART trees for predicting model parameters. For this reason, most recently a larger speech database was analyzed in order to determine the statistically relevant predictor variables for 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 (Rapp, 1998). 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 many applications serves as a reading machine.

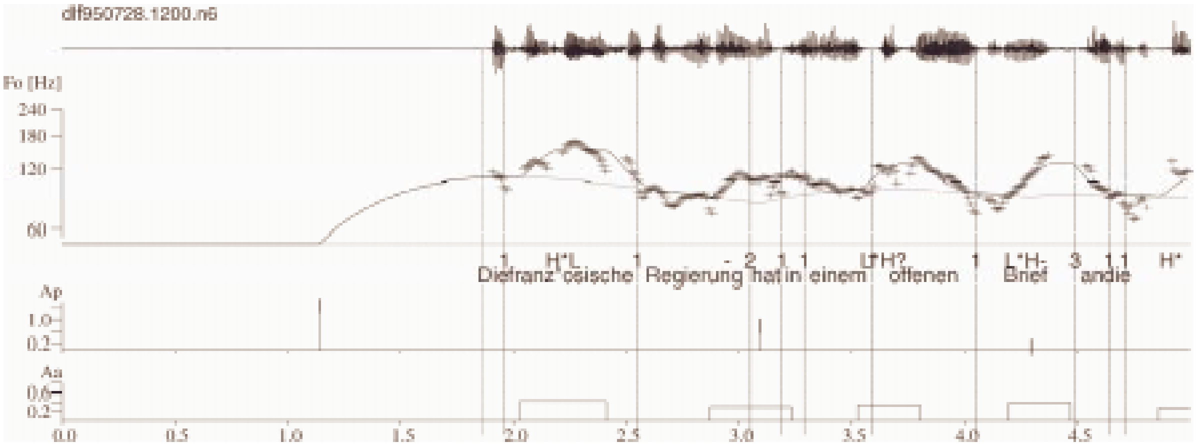

Figure 4. Initial part of an utterance from the database. The figure displays from top to bottom: (1) the speech waveform, (2) the extracted (+-signs) and estimated (solid line) F0 contours, the ToBI labels and text of utterance, the underlying phrase commands (impulses) and accent commands (steps).

The corpus contains boundary labels on the phone, syllable and word levels and linguistic annotations such as part-of-speech. Futhermore, prosodic labels following the Stuttgart G-ToBI system (Mayer, 1995) are provided. The Fujisaki parameters were extracted applying a novel automatic multi-stage approach (Mixdorff, 2000). This method follows the philosophy that not all parts of the F0 contour are equally salient, but are 'highlighted' to a varying degree by the underlying segmental context. Hence F0 modeling in those parts pertaining to accented syllable nuclei (the locations of tone switches) needs to be more accurate than along low-energy voiced consonants in unstressed syllables, for instance.

Results of Analysis

Figure 4 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.

Accentuation

The corpus contains a total number of 13151 syllables. For these a total number of 2931 accent commands were computed. Of these 2400 are aligned with syllables labeled as accented. 177 unaccented syllables preceding prosodic boundaries exhibit an accent command corresponding to a boundary tone $B\uparrow$. A rather small number of 90 accent commands are aligned with accented syllables on their rising as well on their falling slopes, hence forming *hat patterns*.

Alignment. The *information intoneme* $I\downarrow$, and the *non-terminal intoneme* $N\uparrow$ 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 syllable offset time. N-intonemes preceding a prosodic boundary are linked to considerably longer accent commands than those in phrase-initial or medial position. This indicates that the accent command offset is aligned rather with respect to the prosodic boundary than the syllable offset.

A considerable number of accented syllables (N=444) was detected which had not been assigned any accent labels by the human labeller. Figure 4 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.

Prominence. Table 1 shows the relative frequency of accentuation depending on the part-of-speech of the respective word. As could be expected, nouns and proper names are accented more frequently than verbs, which occupy a middle position in the hierarchy, whereas function words such as arcticles and prepositions are very seldom accented. For the categories that are frequently accented the right-most column lists a mean *Aa* reflecting some degree of relative prominence depending on the part-of-speech. As, however, can be seen, differences found in these mean values are little significant. As shown in Wolters & Mixdorff (2000), word prominence is more strongly influenced by the syntactic relationship between words than simply by parts-of-speech.

Table 1. Occurence, frequency of accentuation and mean Aa for a choice of parts-of-speech

Part-of-Speech	Occurence	Accented %	Mean *Aa*
Nouns	1262	75.8	0.28
Proper names	311	78.4	0.32
Adjectives conjugated	333	71.6	0.25
Adjectives non-conjugated	97	85.7	0.28
Past participle of full verbs	172	77.3	0.29
Finite full verbs	227	42.7	0.30
Adverbs	279	41.9	0.29
Conjunctions	115	2.6	-
Finite auxiliary verb	219	3.0	-
Articles	804	1.0	-
Prepositions	621	2.0	-

A very strong factor influencing the *Aa* assigned to a certain word, is whether or not it precedes a deep prosodic boundaries. Pre-boundary accents exhibit a mean *Aa* of ??? against an *Aa* of ??? for phrase-medial accents. The same can be observed for boundary tones (mean *Aa*=???).

Phrasing

All inter-sentence-boundaries were found to be 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 *Ap* 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 useful indicator of boundary strength. In Figure 5 the phrase component extracted for a complete news paragraph is displayed where sentence onsets are marked with arrows. As can be seen from the figure, the magnitudes of the underlying phrase commands nicely reflect the phrasal structure of the paragraph.

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 eight syllables found in Mixdorff (1998) 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.

A Preliminary Model of Syllable Duration

In order to align an F0 contour with the underlying segmental string, F0 model parameters need to be related to the timing grid of an utterance. As was shown for the timing of intonemes in the preceding

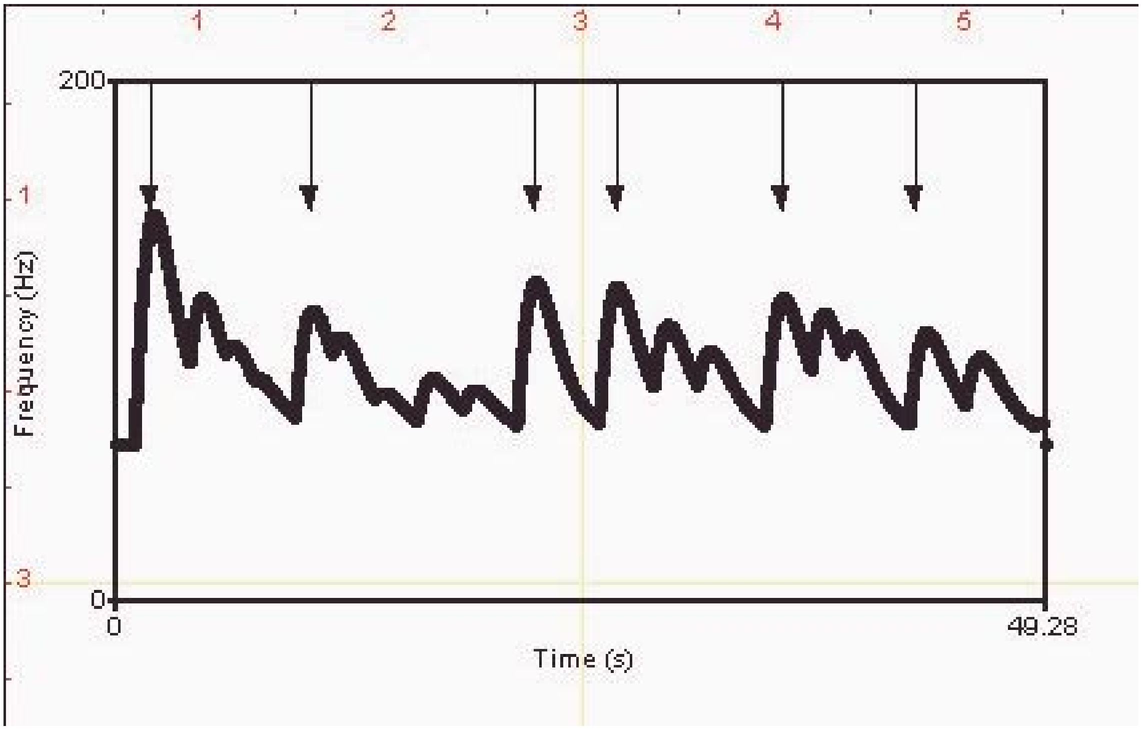

Figure 5. Profile of the phrase component underlying a complete news paragraph. Sentence onsets are marked with vertical arrows.

section, the syllable appears to be an appropriate temporal unit for 'hooking up' F0 movements pertaining to accents.

The timing of tone switches can thus be expressed by relating *T1* and *T2* to syllable onset and offset times, respectively. In a similar fashion, the phrase command onset time *T0* can be related to the onset time of the first syllable in the corresponding phrase, namely by the distance between *T0* and the segmental onset of the phrase.

A regression model of the syllable duration was hence developed which permits to decompose the duration contour into an *intrinsic* part related to the (phonetic) syllable structure and a second, *extrinsic* part related to linguistic factors, such as 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 intrinsic 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{duration_contour} 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 Zellner-Keller (1998), 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.

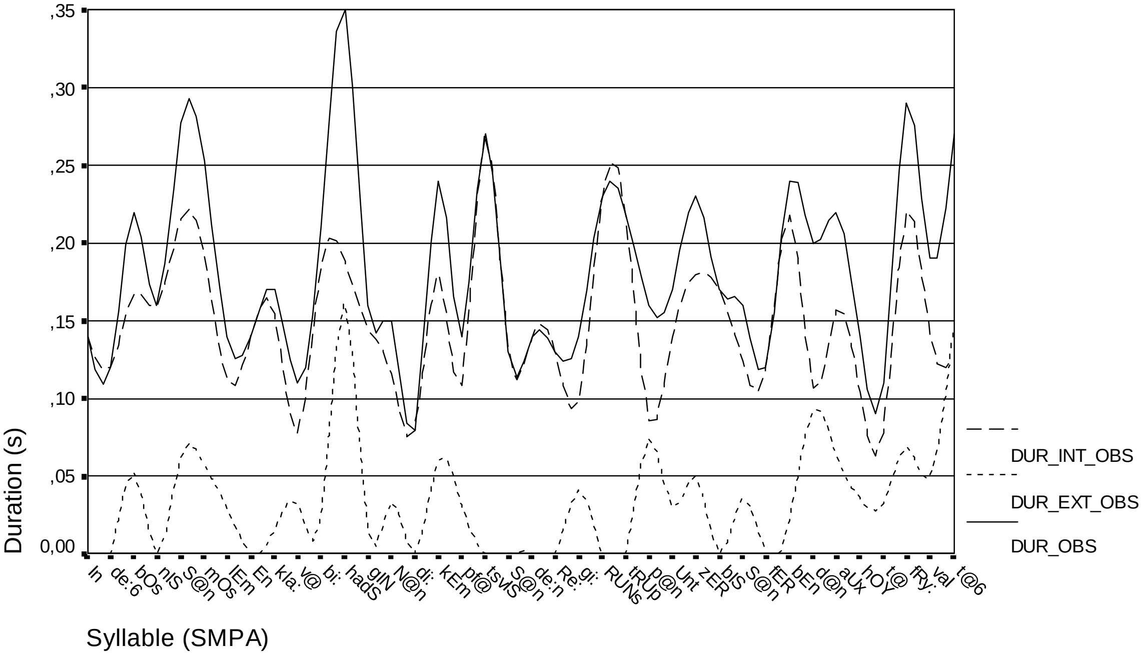

Figure 6. Example of smoothed syllable duration contours for the utterance "In der bosnischen Moslem-Enklave Bihac gingen die Kämpfe zwischen den Regierungstruppen und serbischen Verbänden auch heute früh weiter."-*"In the Bosnian Muslim-enclave Bihac, fights between the government troops and Serbian 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.

Summary and Conclusions

The current paper discussed the linguistically motivated prosody model MFGI which was recently applied to a larger prosodically labeled database. It was shown that model parameters can be readily related to the linguistic information underlying an utterance. Accent commands are typically aligned with accented syllables or syllables bearing boundary tones. Higher level boundaries are marked by the onset of phrase commands whereas the detection of lower level boundaries obviously requires 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.

References

Adriaens, L. (1991). *Ein Modell deutscher Intonation*. Ph.D. thesis, Technical University Eindhoven.

Fujisaki, H. and Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E),* 5(4), 233-241.

Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Fujimura, O. (Ed.). *Vocal Physiology: Voice Production, Mechanisms and Functions* (pp. 347-355). Raven Press Ltd., New York.

Hirschfeld, D. (1996). The Dresden text-to-speech system. In *6th Czech-German Workshop on Speech Processing* (pp. 22-24). Prague, Czech Republic.

Hirst, D., Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline-function. *Travaux de l'Institut de Phonétique d'Aix,* 15, 71-85.

Isačenko, A., & Schädlich, H. (1964). *Untersuchungen über die deutsche Satzintonation*. Akademie-Verlag, Berlin.

Kohler, K. (1991). Studies in German intonation. *Arbeitsberichte Nr. 25, Institut für Phonetik und digitale Sprachverarbeitung*. Universität Kiel.

Mayer, J. (1995). *Transcription of German intonation: The Stuttgart system*. Technischer Bericht, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Mertens, P., d'Alessandro, C. (1995). Pitch contour stylization using a tonal perception model. In *Proceedings of ICPhS* 95, vol. 4 (pp. 228-231). Stockholm, Sweden.

Mixdorff, H. (1998). Intonation patterns of German - model-based quantitative analysis and synthesis of F0 contours. Ph.D thesis TU Dresden, 1998 (http://www.tfh-berlin.de/~mixdorff/thesis.htm).

Mixdorff, H. and Mehnert, D. (1999). Exploring the naturalness of several German high-quality-text-to-speech systems. *Proceedings of Eurospeech '99*, vol.4 (pp.1859-1862). Budapest, Hungary.

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

Möbius, B., Pätzold, M., Hess, W. (1993). Analysis and synthesis of German F0 contours by means of Fujisaki's model. *Speech Communication,* 13, 53-61.

Pierrehumbert, J. (1980). The phonology and phonetics of English intonation. Ph.D. thesis, MIT.

Portele, T., Krämer, J., & Heuft, B. (1995). Parametrisierung von Grundfrequenzkonturen. In *Fortschritte der Akustik - DAGA '95*, Saarbrücken, pp. 991-994.

Rapp, S. (1998). Automatisierte Erstellung von Korpora für die Prosodieforschung, Ph.D thesis Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung.

Stock, E., & Zacharias, C. (1982). *Deutsche Satzintonation*. VEB Verlag Enzyklopädie, Leipzig.

Stöber K.; Portele T.; Wagner P.; Hess W. (1999). Synthesis by word concatenation. *Proceedings of EUROSPEECH '99*., vol. 2, (pp. 619-622). Budapest.

Taylor, P. (1995). The Rise/Fall/Connection Model of Intonation. *Speech Communication,* 15 (1), 169-186.

Wolters, M. & Mixdorff, H. (2000). Evaluating radio news intonation: Autosegmental vs. superpositional modeling. To appear in *Proceedings ICSLP 2000*. Beijing, China.

Zellner-Keller, B. (1998). Prediction of temporal structure for various speech rates. In N. Campbell (Ed.) *Volume on Speech Synthesis*. Springer-Verlag.