Building An Integrated Prosodic Model of German

Hansjörg Mixdorff
Faculty of Computer Sciences

Berlin University of Applied Sciences
mixdorff@tfh-berlin.de

Oliver Jokisch
Laboratory of Acoustics and Speech Communication
Dresden University of Technology
Oliver.Jokisch@ias.et.tu-dresden.de

Abstract

The intellegibility and naturalness of synthetic speech strongly depends on its prosodic quality. Departing from works by Mixdorff on a linguistically motivated model of German intonation based on the Fujisaki model, the current paper presents statistical results concerning the relationship between linguistic and phonetic information underlying an utterance and its prosodic features. Statistical analysis yields, inter alia, the following pairs of *strongest single factor* → *prosodic feature*: boundary depth (right) → syllable duration; boundary depth (left) → phrase command magnitude *Ap*; accent type (intoneme) → accent command amplitude *Aa*. These results were employed for training an FFNN-based integrated prosodic model predicting syllable durations along with syllable-aligned Fujisaki control parameters. Correlations between trained and predicted parameters suggest synergy effects, as they are higher for some parameters than correlations yielded when predicting parameters individually from the same set of input features using a regression model. Informal listening tests with first resynthesis examples showed encouraging results.

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 systems concatenating larger chunks of speech from a data base avoid this problem (see, for instance, [1]), as they preserve the natural prosodic structure at least throughout the chunks chosen and aim to minimize the distortion incurred at the edges, these systems are often domain-specific, and the question of optimal unit-selection still calls for the development of improved prosodic models.

More generally speaking, the production process of prosody and the interrelation between the prosodic features of speech is far from being fully understood.

Earlier work by Mixdorff 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 [4], and proved to produce a high naturalness compared with other approaches [5]. 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 TTS systems for German like DreSS calculate prosodic parameters sequentially, generating syllable durations first and then aligning the *F0* contour appropriately. The modules for predicting duration and *F0* are often developed independently and use features derived from different data sources and environments. This approach 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, and partly explains why synthetic speech is easily identified and rated as being of poor quality. Based on these considerations, the objective of the authors is the development of a prosodic model taking into account the coherence between melodic and rhythmic properties of speech.

2. Properties of the Integrated Model

The model, which is based on the syllable as its basic rhythmic unit, is henceforth to be called an 'integrated prosodic model', as the prosodic parameters (1) syllable duration and (2) *F0* (in terms of Fujisaki control parameters) are predicted from the same data base. A scale factor for syllable intensity is calculated in parallel.

Figure 1 displays the output parameters of the integrated model. For each syllable, the duration and, in the case of accented syllables and syllables bearing boundary tones, the parameters of the accent command assigned to the syllable, are calculated. Along with the amplitude *Aa*, the onset time *T1* and offset time *T2* of the accent command are output, the latter two relative to the onset and offset time of the syllable, respectively.

If a syllable is the first in a prosodic phrase, the onset time *T0* of the phrase command assigned to the phrase is calculated with respect to the onset time of the syllable, as well as the phrase command magnitude *Ap*. The speaker-dependent base frequency *Fb* and time constants *alpha* and *beta* are treated as constants.

Phone duration is calculated from the superordinate syllable's duration taking into account the phone properties found in the database.

In order to capture potential interactions between intonation and rhythm, the prosodic parameters are predicted from a set of linguistic and phonetic input features using a single, feed-forward neural network (FFNN), since calculating syllable durations first and relating *F0* to these in a second step would still result in a sequential model. FFNNs have been shown capable of predicting prosodic parameters directly, as well as in terms of control parameters for the Fujisaki model [6].

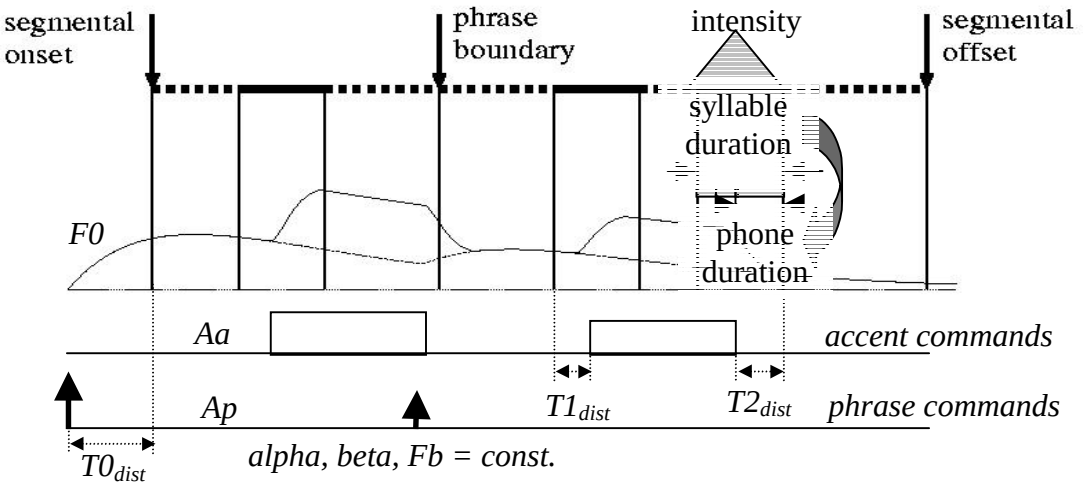

Figure 1. Schematic overview of output parameters of the integrated prosodic model: For each syllable, the duration is calculated (the distance between the vertical lines). Phone duration is derived from syllable duration. Fujisaki control parameters (phrase and accent commands) are described by their amplitudes Ap and Aa, respectively. Command timing is expressed by the relative onset and offset times, $T1_{dist}$ and $T2_{dist}$. The F0 contour (thin solid line) is directly calculated from commands using the Fujisaki formula, keeping alpha, beta and Fb constant for the same speaker. As indicated by the triangle, as a secondary parameter, a scale factor for intensity is output for each syllable.

3. Speech Material and Method of Analysis

A larger speech data base was analysed 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], of a total of 13151 syllables. The decision to use this data base 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.

The corpus contains boundary labels on the phone, syllable and word levels and linguistic annotations such as part-of-speech. The Fujisaki-parameters were extracted applying an automatic multi-stage approach [8]. The mean base frequency *Fb* and time constants *alpha* and *beta* of the current speaker were estimated to be 50.2 Hz, 0.95/s and 20.3/s, respectively.

4. Results of Analysis

Figure 2 shows an example of analysis from the database. The figure displays 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, the underlying phrase and accent commands. Table 1 lists the output parameters of the integrated model and their most important predictor factors along with the correlation coefficients as yielded by statistical analysis. It needs to be noted, that correlations for *Aa*, $T1_{dist}$ and $T2_{dist}$ were calculated only for accented syllables and syllables bearing boundary tones (N=3022), and correlations for *Ap*, $T0_{dist}$ and *pause* (the duration of a pause preceding a prosodic phrase) for syllables which are the first in a prosodic phrase (N=1047). The parameter *strength* indicates whether a syllables is unstressed (0), stressed, but unaccented (1), or stressed and accented (2), i.e. bearing a tone switch. Prosodic boundaries are classified as intra-word / inter-word clitic (depth=0), inter-word (1), inter-phrase (2), inter-sentence (3, at full stops) and inter-paragraph (4, start of news story) boundaries.

From this overview it becomes clear, that the model incorporates information from lower level units (i.e. coda, rhyme, phones) as well as higher levels (word, phrase, sentence, paragraph) in the syllabic parameters. Relationships as rendered by the table are generally in line with the results of earlier works ([3], p.133 ff.). Comparison shows, that, especially in the case of *Aa* and *intensity*, single input variables have relatively little predictive power, whereas for others, such as *syllable duration* and *Ap* single parameters explain more than 40 % of the variance. As expected, a good predictor of *syllable duration* is the sum of mean durations of phone classes (in the data base) pertaining to the syllable, with identical consonant phonemes being treated as different phone classes depending on their position in either coda or rhyme. In the case of *Aa*, the parameter reflecting the relative prominence given to an accented syllable, strong differences were found depending on whether or not an accent precedes

an intra-sentence phrase boundary (mean of *Aa* 0.34 against 0.25, compare example in Figure 2), whereas the part-of-speech of the superordinate word has relatively little influence. The apparently weak contributions of these parameters indicate, that additional information, such as the focal condition (narrow vs. wide focus) associated with an accent, is missing in the data base, as well as a more detailed description of the syntactic environment.

5. Training and Testing the Model

Figure 3 gives an overview of the fully-connected FFNN consisting of 4 layers (24x18x12x8 neurons) using both log and tan-hyperbolic transfer functions. Depending on their ranges, the in- and output parameters are linearly scaled.

Table 2 lists testing results expressed by the correlation between measured and predicted output parameters (center column). It can be seen that correlations for syllable durations are considerably higher than for the Fujisaki control parameters, especially *Aa*. These results confirm the observations in the preceding section concerning possibly missing information in the set of predictor variables. In order to conduct a preliminary test as to whether an integrated prediction of prosodic parameters bears advantages over single-parameter methods, regression models for each output parameter were calculated from the complete set of input features, with correlations given in the right column of Table 2.

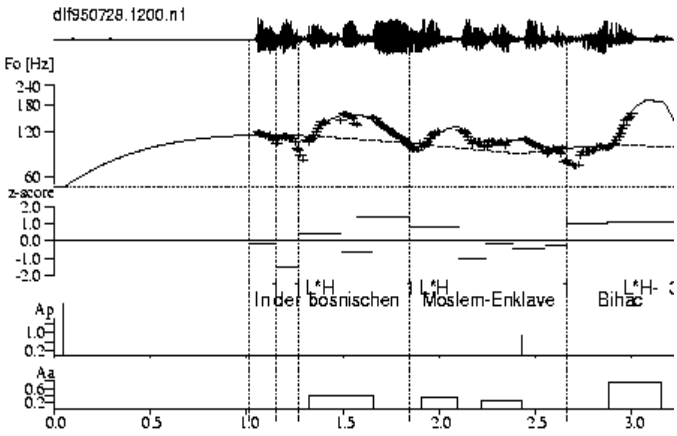

Figure 2:Example of analysis from the data base. From top to bottom: speech waveform, extracted and model-generated F0 contours, duration contour (syllabic z-score), ToBI tier, text of utterance, underlying phrase and accent commands. In the utterance "In der bosnischen Moslem-Enklave Bihac..."-"In the Bosnian Muslim-enclave of Bihac..." the accent command before the phrase boundary at 'Bihac' exhibits a considerably high amplitude Aa.

Table 1: Output parameters and most important predictor variables. t_{on} and t_{off} denote onset and offset time of the current syllable, respectively.

Output Parameter *out* of Model	Predictor Variable *in* of Model	*ρ (out,in)*	*N*
syllable duration	sum of duration means of phone classes in syllable	.640	13151
	boundary depth (right), 0=clitic, 1=word, 2=phrase, 3=sentence, 4=paragraph	.464	13151
	strength (0=unstressed, 1=stressed, 2=accented)	.349	13151
	nucleus schwa/non-schwa	-.191	13151
Aa	type of intoneme (tone switch class)	.257	3022
	part-of-speech	.128	3022
	phrase index in sentence	-.115	3022
$T1_{dist} = T1\text{-}t_{on}$	type of intoneme	.508	3022
	number of phones in syllable onset	.154	3022
$T2_{dist} = T2\text{-}t_{off}$	type of intoneme	.384	3022
	number of phones in syllable rhyme	-.198	3022
Ap	boundary depth (left)	.696	1047
	index of phrase in sentence	-.507	1047
	duration of preceding phrase	.320	1047
	Ap of preceding phrase command	-.184	1047
	duration of current phrase	.110	1047
$T0_{dist} = t_{on}\text{-}T0$	distance from preceding phrase command	.256	1047
intensity (mean frame power *rms* in syllable)	index of phrase in sentence	-.206	13151
	coda voiced	.141	13151
	index of syllable in phrase	-.124	13151
pause	boundary depth (left)	.622	1047
	index of phrase in syllable	-.376	1047

Comparison between FFNN and regression model-based results indicates improvements in five parameters, especially timing parameters and *intensity*, and slight deterioration in the remaining three. This suggests certain synergy effects due to the integrated approach, but obviously requires further investigation.

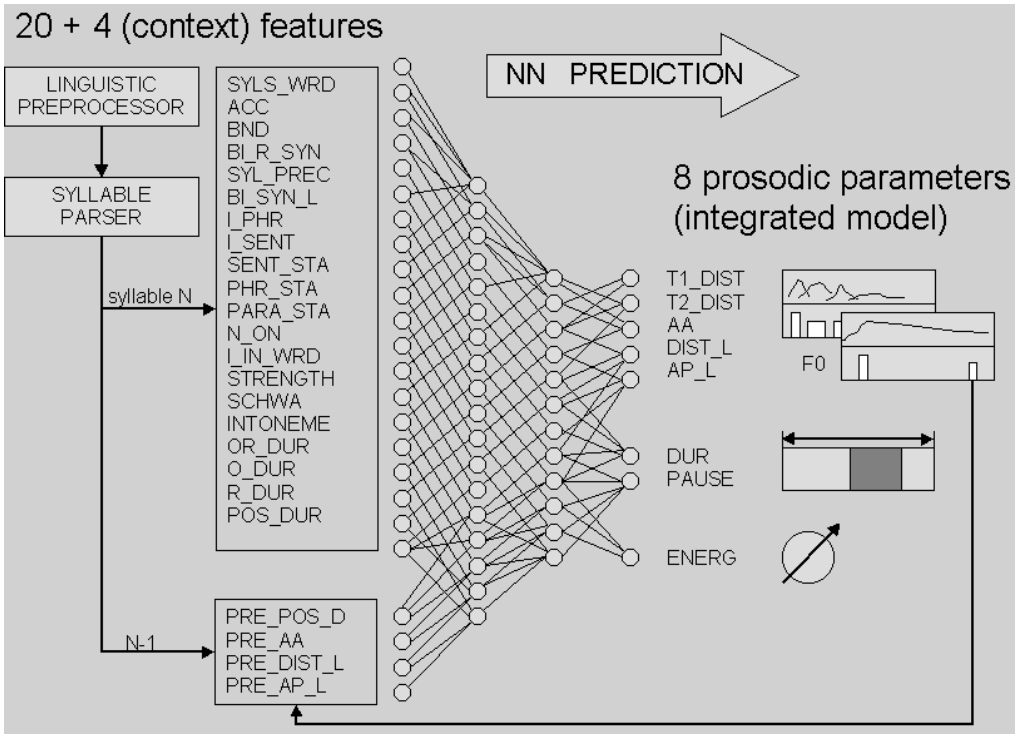

Figure 3:Overview of FFNN-based integrated prosodic model predicting eight parameters from a set of 24 input features.

Table 2:Comparison of prediction results from the FFNN and single parameter regression models. The better correlation values are set in bold type.

out	*ρ (meas.,pred.) FFNN*	*ρ (meas.,pred.) regr. model*	*N*
syll. dur.	**.812**	.809	13151
Aa	.397	**.410**	3022
$T1_{dist}$	**.613**	.578	3022
$T2_{dist}$	**.625**	.587	3022
Ap	.730	**.756**	1047
$T0_{dist}$	**.532**	.349	1047
intensity	**.455**	.373	13151
pause	.725	**.757**	1047

6. Discussion and Conclusions

The current study introduced a novel integrated approach for predicting prosodic features in TTS. The model is based on results of statistical analysis of a larger corpus which were used for training a single FFNN predicting -inter alia- syllable durations along with syllable-aligned Fujisaki control parameters.

Correlations between trained and predicted parameters suggest certain synergy effects, as they are higher for some parameters than correlations yielded when predicting parameters individually from the same set of input features using a regression model. In the case of *intensity* and accent command amplitude *Aa*, obviously additional information needs to be taken into account, for the former supposedly the internal structure of the syllable, and for the latter focal condition and syntactic environment of an accented syllable. Considering the properties of the neighbouring syllables as additional input parameters might as well improve the accuracy of the predictor. Informal listening tests with first resynthesis examples showed encouraging results.

7. 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): 233-241, 1984.

[3] Mixdorff, H., *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), 1998.

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

[5] 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.

[6] Jokisch, O., H. Mixdorff et al., "Learning the parameters of quantitative prosody models ", in *Proceedings ICSLP 2000*, vol.1, pp. 645-648, Beijing, China, 2000.

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

[8] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters", in *Proceedings ICASSP 2000*, vol. 3, 1281-1284, Istanbul, Turkey, 2000.