Comparing a Data-Driven and a Rule-Based Approach to Predicting Prosodic Features of German

Hansjörg Mixdorff [1,2]
[1] Faculty of Computer Sciences
Berlin University of Applied Sciences
Mixdorff@tfh-berlin.de

Oliver Jokisch [2]
[2] Laboratory of Acoustics and Speech Communication
Dresden University of Technology
Oliver.Jokisch@ias.et.tu-dresden.de

Abstract

The perceived quality of synthetic speech strongly depends on its prosodic naturalness. Departing from works by Mixdorff on a linguistically motivated model of German intonation based on the Fujisaki model, the current paper presents a perceptual comparison between a rule-based sequential and a data-driven integrated model of prosody. The experiment comprised resynthesis and diphone synthesis stimuli produced with prosodic features predicted by the two models. In addition, a number of reference stimuli from an earlier experiment were included which were produced by means of controlled prosodic degradation. Isolated sentences from two different corpora served as speech material. Results show that the integrated model outperforms the rule-based model in terms of the accuracy of phone durations predicted. In terms of the *F0* contour generation, however, the integrated model is not rated better. Furthermore, these findings are only significant in the case of resynthesis stimuli whereas diphone synthesis stimuli are judged generally as being of poor quality, irrespective of the prosodic model. This fact clearly speaks against testing diphone synthesis samples against resynthesis in the same evaluation, as the segmental quality overrules the prosodic quality.

1. Introduction

It is widely acknowledged that the intelligibility and perceived naturalness of synthetic speech strongly depends on the prosodic quality. Recent systems concatenating larger chunks of speech from a data base achieve a considerably high quality (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, however, are often domain-specific, and the question of optimal unit-selection still calls for the development of improved prosodic models.

Earlier work by Mixdorff focussed on a model of German intonation which uses the quantitative Fujisaki formulation of the production process of *F0* [2] for parameterizing *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 shortcomings 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 current paper presents results from an experiment comparing the rule-based prosodic model (henceforth: 'RBM') from [4] with a novel integrated approach (henceforth: 'IGM'). The integrated model parameters are predicted with a four-layer feed-forward neural network (FFNN) introduced in an earlier work by Jokisch [5].

The approaches differ in several respects:

- The RBM calculates phone durations based on a Klatt/Kohler-style formula.
- According to the phone durations determined, the *F0* contour is aligned with nuclear vowels, based on a set of heuristic rules ([3], p. 238 ff.) for determining the underlying Fujisaki parameters. These rules were established by analysis of a small corpus and optimized with respect to the TTS system DRESS.
- The IGM conjunctly calculates syllable durations and syllable-aligned Fujisaki parameters using an FFNN trained on data from a larger, prosodically labeled speech corpus
- phone durations are calculated with respect to the superordinate syllable duration.

The two models will be discussed in more detail in the following sections.

2. The Rule-based Model

In the original model, phone duration is calculated using a Klatt/Kohler-style formula in which departing from a phone-specific inherent duration, by subsequent application of rules, factors, such as the proximity of phrase boundaries, the phonetic properties of adjacent phones or the presence/absence of lexical stress are taken into account. If a phone is part of a function word, for instance, it is compressed, but not below a phone-specific minimum value.
Based on the predicted phone durations, following a set of heuristic rules, Fujisaki model commands are aligned with the segmental string. Accent commands are set up according to the assignment of tone switches to accented or pre-boundary syllables (boundary tones). The fine alignment of accent command is realized with the onset of the nuclear vowel of an accented syllable as a reference point. In the case of rising accents, the onset of an accent command is aligned with this reference, and in the case of falling accents, the offset of an accent command. The respective offset and onset times are chosen considering the offset/onset of following/preceding nuclear vowels and a minimum criterion for accent command duration. In the case of accents close to utterance-medial boundaries, the boundary itself determines the accent command offset time. The amplitude *Aa* is set with respect to the type of intoneme. Phrase commands are aligned with phrase boundaries, typically by setting up the phrase command at a distance of 1/*alpha* before the segmental onset of the phrase, for the phrase component to reach its maximum at the segmental onset. The phrase command magnitude *Ap* is set depending on the boundary depth and the number of syllables in the preceding phrase.

3. The Integrated Model

The model, first introduced in [6], predicts the prosodic parameters (1) syllable duration, (2) *F0* (in terms of Fujisaki control parameters), (3) pause duration, and (4) syllable energy based on prosodic labels learned from a data base. The underlying corpus is part of a 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 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. Table 1 lists the output parameters of the IGM which treats the syllable as its basic rhythmic unit.
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 defined with respect to the onset time of the syllable, and calculated together with the magnitude *Ap* of the phrase command. 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 FFNN, since calculating syllable durations first and relating *F0* to these in a second step would still result in a sequential model.

Table 1: Output parameters of the IGM. t_{on} and t_{off} denote onset and offset time of the current syllable, respectively. The parameters alpha, beta and Fb are assumed to be constant for the same speaker. The right column lists the correlation between measured and predicted output parameters.

Output Parameter of Model	Calculated as	N of tokens in data base	Correlation between predicted and measured parameters
syllable duration	t_{off} - t_{on}	13151	.812
Aa	-	3022	.397
$T1_{dist}$	$T1$-t_{on}	3022	.613
$T2_{dist}$	$T2$-t_{off}	3022	.625
Ap	-	1047	.730
$T0_{dist}$	t_{on}-$T0$	1047	.532
energy	mean frame power *rms* in syllable	13151	.455
pause	inter-phrase pause duration	1047	.725

3.1. Results of Analysis

Statistical analysis was performed in order to determine the linguistic and phonetic factors with the strongest influence on the output parameters of the model given in Table 1. It should be noted that predictor factors for *Aa*, $T1_{dist}$ and $T2_{dist}$ were determined only for accented syllables and syllables bearing boundary tones (N=3022), and factors influencing *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).
Analysis shows that in the case of *syllable duration*, the depth of the prosodic boundary to the right, 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), is the strongest **extrinsic**[1] predictor factor (ρ=.464), followed by the factor *strength* which indicates whether a syllable is unstressed (0),

[1] The term *extrinsic* denotes factors not pertaining to the syllable structure as opposed to *intrinsic* factors, i.e. the properties of the phones in the syllable.

stressed, but unaccented (1), or stressed and accented (2), i.e. bearing a tone switch (ρ=.349). The best **intrinsic** factor for predicting *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 onset or rhyme (ρ=.640). It becomes clear that the integrated prosodic model incorporates information from lower level units (i.e. onset, rhyme, phones) as well as higher levels (word, phrase, sentence, paragraph) in the syllabic parameters.

In the case of *Aa,* the parameter reflecting the relative prominence lent 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). The type of accent (non-terminal phrase-final, non-terminal phrase-medial, declarative final) is therefore the most important predictor factor for Aa (ρ=.257) whereas the part-of-speech of the superordinate word has relatively little influence (ρ=.128). The apparently weak contributions of these parameters, also reflected by the low correlation of 0.397 between measured and predicted *Aa* (right column in Table 1) indicate that additional information associated with an accent is missing in the data base. This problem will be addressed in more detail in the following section.

4. Predicting Prominence

As indicated, prominence in terms of accent command amplitude *Aa* assigned to constituents in an utterance can only be predicted very coarsely from input information such as the part-of-speech and the type and position of an accent. Although generally speaking words can be roughly classified as content or function words, with the latter being accented only in contrastive contexts, assigning prominence to content words proves to be a difficult task. As shown in Table 2 which displays the

Table 2: Selection of frequent parts-of-speech with frequency of accentuation and mean accent command amplitde Aa for accented cases.

Part-of-Speech	**Occurence**	**Accented %**	**Mean Aa**
Nouns	1262	75.8	0.28
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	
Possessive pronouns	65	3.0	
Personal Pronouns	83	2.4	
Articles	804	1.0	
Prepositions	621	2.0	

percentage of accentuation and mean *Aa* for a selection of parts-of-speech, verbs, especially in sentence-final position are less often accented than nouns. The average accent command amplitude assigned to accented verbs, however, does not significantly differ from that of nouns. This indicates that for predicting *Aa* of a content word other factors need to be taken into account, such as the

linguistic (syntactic and semantic) context and pragmatic requirements. For this reason and in addition to the gross statistical evaluation of the entire corpus, a phrase-wise analysis was performed on half of the corpus. In the following, a small number of accentuation patterns will be discussed that could be identified stably in the corpus. Instances of these patterns, however, are rather infrequent (N < 100) and therefore statistically weak compared with the number of 3022 accented syllables in the corpus.

Enumerations. Examining instances of lists containing three items, typically names, ("A, B and C...") yields higher prominence for the first and the third item than for the second, as in the examples shown in the following table (*Aa* given in brackets):

Laupheim, (.39)	Peisenberg (.22)	und Speyer (.35)
Bosnien, (.29)	Kroatien (.23)	und Serbien (.29)
Münch, (.50)	Perschau (.33)	und Schreiber (.46)
Deutschland (.36)	Frankreich (.26)	und die Niederlande (.32)

Sequence of Function and Name. The news stories very often refer to persons of public interest who are introduced with their function and name. In these constructs, the function is generally less prominent than the name. The following table renders a few examples (*Aa* given in brackets).

Bundesaußenminister (.09) Kinkel (.33)
der frühere(.24) Regierungschef (.00) Münch (.68)
Sachsen-Anhalts(.12) Regierungschef (.00) Höppner (.44)
der bosnische (.13) Außenminister (.00) Čečebej (.50)

Given and New. In this context 'Given and New' simply refers to whether or not a word (typically a name) has already been mentioned in the current news story. Consistent decrease of prominence can mostly be observed between first and second mention, especially when they occur in consecutive sentences. If the distance is larger, as for instance in a third mention, the word prominence is likely to increase again. The following table renders a few examples (*Aa* for first and second mention given in brackets).

Carter (.32, .11)	Horstmann (.52, .20)	Scharping (.36, .22)
Burns (.59, .13)	Castro (.66, .24)	Masowiecki (.53, .36)
Rau (.82, .29)	Scharping (.35, .16)	Däubler-Gmelin (.37, .13)

This pattern only applies to repeated mentions of the same word. If a person is introduced by his name and later referred to by his function, a similar decrease of prominence is generally not observed.

The small selection of recurring accentuation patterns presented in the current section suggests possible limitations of a statistical approach to predicting prominences. Since instances of these patterns occur infrequently in the data base, there might be too few for influencing the behaviour of the neural network. Furthermore, this kind of infrequent, yet stable phenomena could more easily be taken care of by formulating prominence rules (for lists, for patterns of function+name, for second mentions etc.).

5. The Perceptual Evaluation

5.1. Experiment Design

A series of perception experiments for evaluating the quality of the IGM was designed. Twelve isolated sentences of varying complexity (between 13 and 44 syllables) were chosen and resynthesized using predicted prosodic parameters, i.e. syllable durations and *F0* contours. All stimuli were created by applying the PSOLA resynthesis functionality of the software *PRAAT* (© P.Boersma/D. Weenink) and replacing *DurationTiers* and *PitchTiers.*

In the first experiment, reported in [9], a reference matrix of stimuli was created by controlled degrading of the prosodic features of the natural utterances. The degree of degradation was determined by the cross-correlation between natural and modified/predicted parameters, i.e. rho_{dur} for durations and rho_{F0} for *F0*. The IGM was rated better than 'degraded stimuli' of comparable rho_{dur} and rho_{F0}.

The second experiment reported here aimed at comparing the original sequential RBM evaluated in [4] with the integrated approach. For the stimuli generated using the IGM, an average rho_{F0} of 0.55 was calculated, and an average rho_{dur} of 0.82. For the RBM, an average rho_{F0} of 0.54 was calculated, and an average rho_{dur} of 0.71.

Six sentences from the corpus used in [4] were selected (henceforth referred to as 'J-Set') and six from the Stuttgart data base sentences already used in the first experiment ('D-Set'). In addition, a selection of diphone synthesis stimuli was created using MBROLA, using the prosodic information for the resynthesis conditions. As a reference, originals and a selection of 'degraded stimuli' from the first experiment were also included in the evaluation. The complete selection of stimuli is displayed in Figure 1.
The J-Set was mainly chosen for examining the performance of the integrated model on sentences not pertaining to the newsreading corpus.
Subjects taking part in the experiment were 21 students of Telecommunication Engineering at Berlin University of Applied Sciences in their fourth year. They were informed that the experiment dealt with the quality of synthetic speech, but not about the details of parameters manipulated. The experiment took about 60 minutes to perform, with a ten-minute break after the first half of the stimuli.
Of each sentence, 16 versions were created, yielding 192 different stimuli. In order to test the consistency of the quality judgement, the stimuli from the prosodic models to be compared were included twice, bringing the total number of stimuli to 264. The subjects were provided with forms and requested to assess the quality of the stimuli with grades between 1 (very good) and 5 (very bad), according to the German grading system. Intermediate grades 2, 3 and 4 were explained as corresponding to judgments of 'good', 'acceptable' and 'bad'.
The sequence of the first sixteen samples in the first turn was presented twice, in order to familarize the subjects with the 'quality spectrum' of the stimuli. During the assessment phase of the experiment, stimuli were presented in randomized order and played back twice for every decision, while observing that consecutive stimuli pertained to different sentences.

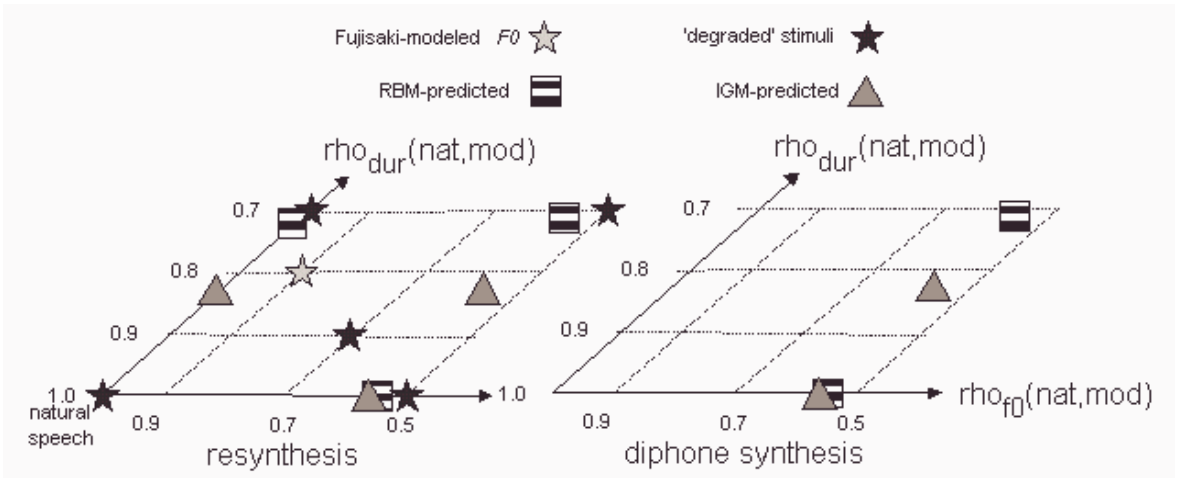

Figure 1: Arrays of stimuli used in the perception experiment. The dark-grey triangles denote versions using IGM-based prediction, striped rectangles versions from RBM prediction. The black stars denote reference stimuli created by prosodic degrading, the grey star a version produced using extracted Fujisaki-parameter configurations.

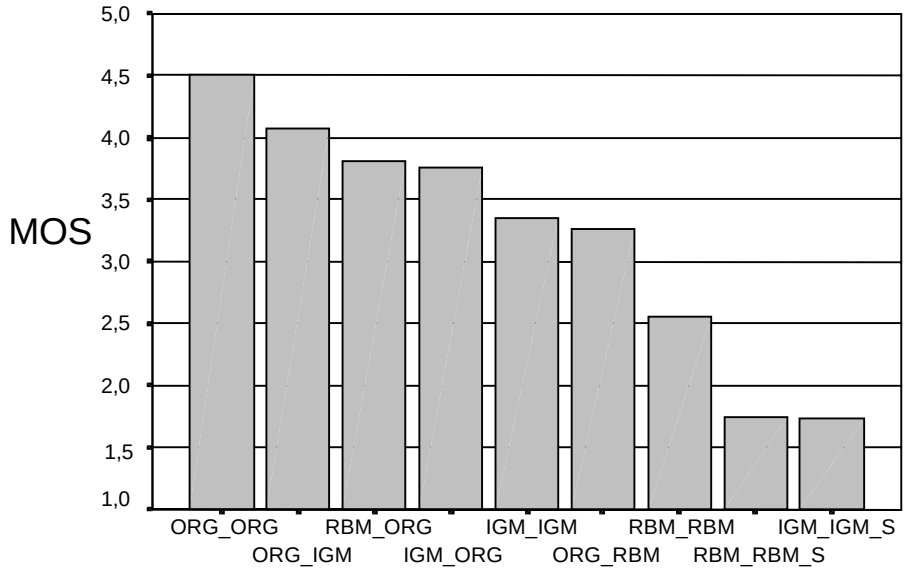

Figure 2: Bar chart of approach-wise mean opinion scores. Approaches are labeled as follows: `f0-algorithm_duration-algorithm[_S]`. `ORG` *stands for original prosodic parameters,* `IGM` *for integrated model-based, and* `RBM` for *rule-based model-derived parameters.* `_S` *denotes approaches using diphone synthesis. As can be seen, the integrated model outperforms the rule-based model mainly by its better duration prediction.*

5.2. Results

In order to yield a score increasing with perceived stimulus quality (*MOS*), grades assigned by the subjects were subtracted from a value of 6, producing a scale from 1 to 5. The *MOS* averaged over all sentences depending on the approach is displayed in Figure 2. As can be seen, the natural speech

stimuli were not unanimously rated 'very good' and only reach a *MOS* of 4.51, with an average of 59% of the 'original' stimuli in the set of utterances being assigned 'grade 1'.
The second best rating of 4.08 was assigned to stimuli with durations produced with the IGM and natural *F0*. The corresponding value for the RBM is 3.27. A similar distance (3.36 vs. 2.56) is found for stimuli where durations as well as *F0* contours were predicted by the models. Comparison with stimuli using natural durations and model-based *F0* (3.76 vs. 3.81), however, indicates that the quality difference between the models must be attributed to the better duration prediction of the IGM, whereas the F0 contours from the RBM are judged even slightly better. Differences between the prosodic models are completely leveled in the case of diphone based stimuli (1.74 vs. 1.75) which were unanimously placed at the lower end of the quality spectrum. The performance of the IGM on the J-Set was not found to be significantly different from that on the D-Set. Although all subjects actually made use of all available grades from 1 to 5, the individual averages vary between 2.44 and 3.56 (mean=2.88, slightly less than the scale means of 3.0). The mean inter-subject correlation was found to be of 0.611, indicating considerable individual variations.

6. Discussion and Conclusions

The current study compared the performance of a rule-based sequential and a data-driven integrated prosodic model of German. Perceptual evaluation indicates that the IGM outperforms the rule-based model by the quality of segment durations. However, statistical analysis also shows that features currently derived from plain text are not sufficient for accurately predicting prominences of accented items in an utterance. For this reason, the IGM does not gain any improvement over the RBM in terms of the *F0* contours predicted. As discussed in section 4, prominence is influenced by a number of additional infrequent, but stable factors. The current prediction which is simply based on position, accent type and part-of-speech captures only a fraction of the necessary information. It also needs to be questioned whether statistical approaches, currently being ‘state-of-the-art’, are capable of modeling infrequent events, and how they could be complemented by a set of rules. The experiment outcome shows that presenting resynthesis and diphone synthesis stimuli in the same experiment automatically places the latter on the lower end of the quality spectrum where segmental quality overrules prosodic quality.

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, Hungary 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] 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] Jokisch, O., Mixdorff, H., Kruschke, H., Kordon, U., “Learning the parameters of quantitative prosody models”, in *Proceedings ICSLP 2000*, 645-648, Beijing, 2000.

[6] Mixdorff, H. and Jokisch, O., “Building an Integrated Prosodic Model of German.” Accepted for presentation at *Eurospeech 2001*, Aalborg, Denmark, 2001.

[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.

[9] Mixdorff, H. and Jokisch, O., “Implementing and Evaluating an Integrated Approach to Modeling German Prosody”. Accepted for presentation at the 4th ISCA Workshop on Speech Synthesis. Perthshire, Scotland, 2001.