INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 45–55, 2003

Evaluating the Quality of an Integrated Model of German Prosody

Disk
followed

HANSJÖRG MIXDORFF
Faculty of Computer Sciences, Berlin University of Applied Sciences, Germany; Laboratory of Acoustics and Speech Communication, Dresden University of Technology, Germany
mixdorff@tfh-berlin.de

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

Abstract. The perceived quality of synthetic speech strongly depends on its prosodic naturalness. Departing from earlier works by Mixdorff on a linguistically motivated model of German intonation based on the Fujisaki model, an integrated approach to predicting $F0$ along with syllable duration and energy was developed. The current paper first presents some statistical results concerning the relationship between linguistic and phonetic information underlying an utterance and its prosodic features. These results were employed for training the MFN-based integrated prosodic model predicting syllable duration and energy along with syllable-aligned Fujisaki control parameters. The paper then focusses on the method of perceptual evaluation developed, comparing resynthesis stimuli created by controlled prosodic degrading of natural speech with stimuli created using the integrated model. The results indicate that the integrated model generally receives better ratings than degraded stimuli with comparable durational and $F0$ deviations from the original. An important outcome is the observation that the accuracy of the predicted syllable durations appears to be a stronger factor with respect to the perceived quality than the accuracy of the predicted $F0$ contour.

Keywords: prosody in text- to-speech, Fujisaki model, evaluation of 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 database achieve a considerably high quality (see, for instance, Stöber et al., 1999), 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 $F0$ (Fujisaki and Hirose, 1984) 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 (Mixdorff, 1998). The model was integrated into the TU Dresden TTS system DreSS (Hoffmann, 1999), and proved to produce a high naturalness compared with other approaches (Mixdorff and Mehnert, 1999). 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. This method does not sufficiently take into account 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. In other words, the modeling of the production process of prosody and the interrelations between the prosodic features of speech is far from being a solved problem. 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.

The model is henceforth to be called an 'integrated prosodic model' (Mixdorff and Jokisch, 2001), as the prosodic parameters (1) syllable duration, (2) $F0$ (in terms of Fujisaki control parameters), (3) pause duration, and (4) syllable energy, are predicted from the same database.

2. Speech Material and Method of Analysis

A speech database 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 (Rapp, 1998), of a total of 13.151 syllables.

The corpus (henceforth referred to as the 'D 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 (Mixdorff, 2000). The mean base frequency *Fb*, and time constants α and β of the speaker of the D corpus were estimated to be 50.2 Hz, 0.95/s and 20/s, respectively.

The phone boundaries had been determined by means of forced alignment. Based on this segmentation, mean duration and standard deviation was calculated for all phones, with identical consonant phonemes being treated as different phone classes depending on their position in either onset or coda.

10.000 syllables of the D corpus were later used for training, and the remaining 3.151 syllables for numerically testing the integrated model (see Section 5).

Twelve sentences were randomly chosen from the D corpus for perceptual evaluation (henceforth called the 'D set', see Section 6).

In order to evaluate the performance of the model on a different speaker and a different speaking style (see Section 6.3), six utterances from a corpus of simple one and two-clause sentences (henceforth the 'J set') which had been used in an earlier study (Mixdorff and Mehnert, 1999) were chosen. These were labeled and analyzed in the same way as the D corpus. *Fb* and time constants α and β of the speaker of the J set were estimated to be 75.0 Hz, 2.0/s and 20.0/s, respectively.

3. Properties of the Integrated Model

Table 1 lists the output parameters of the integrated model 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 .

The speaker-dependent base frequency *Fb* and time constants α and β are treated as constants.

Table 1. Output parameters of the integrated prosodic model. t_{on} and t_{off} denote onset and offset time of the current syllable, respectively. The parameter α, β and *Fb* are assumed to be constant for the same speaker.

Output parameter of model	Calculated as	*N* of tokens in database
Syllable duration	$t_{\mathrm{off}} - t_{\mathrm{on}}$	13.151
Aa	–	3.022
$T1_{\mathrm{dist}}$	$T1 - t_{\mathrm{on}}$	3.022
$T2_{\mathrm{dist}}$	$T2 - t_{\mathrm{off}}$	3.022
Ap	–	1.047
$T0_{\mathrm{dist}}$	$t_{\mathrm{on}} - T0$	1.047
Energy	Mean frame power rms in syllable	13.151
Pause	Inter-phrase pause duration	1.047

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

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, multi-layer feed-forward neural network (MFN), since calculating syllable durations first and relating $F0$ to these in a second step would still result in a sequential model. MFNs have been shown capable of predicting prosodic parameters directly, as well as in terms of control parameters for the Fujisaki model (Jokisch et al., 2000).

4. Results of Analysis

Statistical analysis was performed on the D corpus in order to determine the linguistic and phonetic factors with the strongest influence on the output parameters of the model given in Table 1. The selection of parameters was based on earlier works by the first author concerning the relationship between linguistic factors and their influence on the Fujisaki parameters (Mixdorff, 1998). These relationships were reinvestigated on the D corpus by means of calculating correlations between potential predictor factors and predicted parameters, and by performing factor analysis. A main rationale in this context was that the neural net was viewed as a tool for jointly predicting the prosodic parameters from an already established set of input parameters, but not for examining the respective relevance of these parameters by simply throwing everything at the network. Furthermore, only those input parameters were taken into account which were potentially available from a TTS front-end processing plain text.

It should be noted that predictor factors for Aa, $T1_{\text{dist}}$ and $T2_{\text{dist}}$ were determined only for accented syllables and syllables bearing boundary tones ($N = 3.022$), and factors influencing Ap, $T0_{\text{dist}}$ and *pause* (the duration of a pause preceding a prosodic phrase) for syllables which are the first in a prosodic phrase ($N = 1.047$).

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 ($\rho = 0.464$), followed by the factor *strength* which indicates whether a syllable is unstressed (0), stressed, but unaccented (1), or stressed and accented (2), i.e. bearing a tone switch ($\rho = 0.349$). As expected, the best **intrinsic** factor for predicting *syllable duration* is the sum of mean durations of phone classes (in the D corpus) pertaining to the syllable, ($\rho = 0.640$).

Relationships established between linguistic/phonetic factors and Fujisaki control parameter are generally in line with the results of earlier works (Mixdorff, 1998, p. 133 ff.).

Comparison shows, that, especially in the case of *Aa* and *energy*, individual factors have relatively little predictive power, whereas for others, such as *syllable duration* and *Ap* single parameters explain more than 40% of the variance.

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). The type of accent (non-terminal phrase-final, non-terminal phrase-medial, declarative final) is therefore the most important predictor factor for *Aa* ($\rho = 0.257$) whereas the part-of-speech of the superordinate word has relatively little influence ($\rho = 0.128$). 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 labels of the D corpus, as well as a more detailed description of the syntactic environment.

The following is a complete list of the twenty input parameters of the model:

Syllable level parameters
- Sum of mean durations of phones in syllable
- Sum of mean durations of phones in onset
- Sum of mean durations of phones in rhyme
- Nuclear vowel schwa/non-schwa
- Number of phones in onset

Word level parameters
- Index of syllable in word
- Part-of-speech of word (32 duration classes)
- Number of syllables in word
- Lexical word accent (0/1)

Features on the phrase level and above
- Syllables in preceding phrase
- Boundary tone (0/1, before phrase boundaries)
- Break index to the left (0–4)
- Break index to the right (0–4)

(*Continued on next page.*)

(*Continued*).

Index of phrase in sentence
Index of sentence in paragraph
Start of phrase (0/1)
Start of paragraph (0/1)
Start of sentence (0/1)
Type of accent ('intoneme,' three classes)
Syllable strength (0–2)

It becomes clear, that the integrated prosodic 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.

5. Training and Testing the Model

Based on the results of analysis discussed in the preceding section, twenty syllabic features were selected as a syllable-based input vector, and augmented by four context parameters: (for accented syllables) the part-of-speech of the preceding accented word and the amplitude *Aa* assigned to the preceding accent command, and, in the case of phrase-initial syllables, the properties of the preceding phrase commands ($T0_{\text{dist}}$, *Ap*). The output vector consists of five Fujisaki model parameters with relative timing controlling the $F0$ *contour* ($T1_{\text{dis}}$, $T2_{\text{dis}}$, *Aa*, $T0_{\text{dist}}$, *Ap*), the current syllable duration and the duration of a potential pre-syllabic pause (*syllable duration*, *pause*), and also a signal energy parameter (*energy*).

The training and prediction tasks are solved by a multi-layer feed-forward neural network (MFN) of four layers (24 × 18 × 12 × 8 neurons). According to Kolmogorov's Mapping Neural Network Existence Theorem a three-layer network in theory would be sufficient for the mentioned task. From a practical point of view, however, this theorem does not provide a systematic method for determining the appropriate topology of the three-layer network and the number of neurons in the hidden layer may be extremely high. Networks including more hidden layers may train faster and can better generalize on certain problem classes (compare also Hecht-Nielsen, 1990, 122ff., and Zell, 1994, 558ff.).

In the case of this study the adequate size of the two hidden layers was iteratively determined by adding or deleting single neurons in both layers and comparing the converging error criteria on the test set. The systematic *trial and error* method seems to be appropriate considering the complex network I/O vectors. Certainly, other approaches like *local experts* or evolutionary strategies including *pruning* can be also used to find the adequate topology. Indeed, in this study, the available database size (13.151 syllables) is a more critical fact and comparably small with regard to the *rule of thumb* requiring about 10 times more linearly independent sets than weights in the network. Nevertheless, the resulting topology apparently does not cause any convergence problems or over-learning.

Depending on the parameter ranges the input and output parameters are linearly scaled. Both, log and tan-hyperbolic transfer functions are used. The NN is trained using a teaching input (D corpus) and standard error backpropagation, minimizing the root mean square error (RMSE) between teaching input and net output.

The D corpus was subdivided into a training set (10.000 syllables) and an independent test set (3.151 syllables). Observing the RMSE in the test set an over-adaptation to the training data was avoided even at a total of 500–1.500 training cycles. Although the network has a fairly simple structure, it is apparently suited to predict the eight output parameters of the integrated prosodic model concurrently. The accuracy of prediction for individual parameters obviously depends on the predictive power of the 24 selected input parameters and the quality of the prosodic labels in the database from which they are computed (see Section 4). Table 2 shows the RMSE, means and standard deviations of trained and predicted output parameters.

It becomes clear that the predicted parameters exhibit reduced standard deviations, indicating the averaging behavior of the neural network. After rescaling all parameters, and relating the relative timing parameters of the Fujisaki model to the timing of the

Table 2. RMSE between trained and predicted output parameters and their respective means and standard deviations.

Parameter	Overall RMSE	Mean (trained)	s.d. (trained)	Mean (pred.)	s.d. (pred.)
$T1_{\text{dist}}$	0.058 s	0.037 s	0.152 s	0.051 s	0.097 s
$T2_{\text{dist}}$	0.068 s	0.039 s	0.181 s	0.052 s	0.118 s
Aa	0.059	0.293	0.165	0.273	0.064
$T0_{\text{dist}}$	0.138 s	0.435 s	0.544 s	0.432 s	0.196 s
Ap	0.140	1.10	0.62	1.08	0.57
Syllab.dur.	0.046 s	0.189 s	0.078 s	0.191 s	0.064 s
Pause	0.069 s	0.290 s	0.348 s	0.290	0.255 s
Energy	689.8	1697.0	774.5	1677.2	360.4

underlying syllable chain, the model commands can be fed into the Fujisaki model, producing a time-aligned $F0$ contour.

The authors consider the combination of a data-driven approach, such as a neural network, with a rule-based prosodic model a hybrid architecture since the overall behavior of the model differs from conventional, strictly rule-based or daten-driven models. Furthermore, neither feature extraction nor training can be simply split into an either rule-based or a data-driven part. This definition is in line with the work of Corrigan et al. (1997) and the *Hybrid Data-Driven Architecture (HYDRA* Jokisch et al., 1998). This architecture has been successfully exploited for the rapid adaptation to prosodic model parameters using a well-defined rule-based core (Jokisch et al., 2000).

6. Evaluating the Model

6.1. Experiment Design

A series of perception experiments for evaluating the quality of the prosodic model was designed. Twelve sentences randomly chosen from the D corpus of varying complexity (between 13 and 44 syllables, henceforth referred to as the 'D set') were resynthesized using predicted prosodic parameters, i.e. syllable durations and $F0$ contours. As a reference matrix and in order to examine the relative contributions of durational and intonational quality to the overall judgment, an array 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. In the case of syllable durations, degradation was achieved by incremental compressing or stretching, yielding overall target cross-correlations rho_{dur} of 0.9, 0.8, 0.7, and 0.6. These target correlations were chosen on the basis of preliminary tests which had also indicated that correlations for the $F0$ contours needed to be considerably lower in order to yield comparable effects. All stimuli were created by applying the PSOLA resynthesis functionality of the software *PRAAT* (© P. Boersma/D. Weenink) and replacing *DurationTiers* and *PitchTiers.*

The degrading of the $F0$ contours was not performed on the individual $F0$ values, but by modifying the Fujisaki control parameters estimated from the natural utterances. Since microprosody is absent in the smoothed $F0$ contours produced by the Fujisaki model, the average cross-correlation rho_{F0} between original and modeled contours is of 0.93, measured for frames of 10 ms. Departing from the original Fujisaki parameter configurations, by incremental reduction or increase of parameter values, $F0$ contours were created with cross-correlation coefficients rho_{F0} of 0.7, 0.5 and 0.3, while observing that Aa and Ap cannot assume negative values (see Fig. 1 for examples of original and degraded $F0$ contours).

In order to equally distribute the error onto all control parameters involved ($T1_{\text{dis}}$, $T2_{\text{dis}}$, Aa, $T0_{\text{dist}}$, Ap), the increment was adjusted individually for each parameter until the parameter-based cross-correlation between original and modified parameters was nearly equal, while observing the target value for the $F0$ contour-based cross-correlation.

In this context it must be noted that, since the $F0$ contour is computed using alignment information based on syllabic durations, manipulations in the durational domain cause time-warping in the $F0$ contour and hence further degradation. In the current study we saw two ways of realizing the $F0$ modifications in the degraded stimuli: either (1) to adjust the timing of the phrase and accent commands with respect to the time-warped syllables, keeping α and β constant, or (2) time-warp the $F0$ contour calculated for the original syllable durations, preserving its 'vertical warp'. Since option (1) resulted in melodically different contours depending on the modification of the syllable durations, we chose option (2) which slightly changes the slopes of the $F0$ contour and therefore the effective α and β, but preserves the vertical excursions of the $F0$ contour regardless of the syllabic modification.

Subjects taking part in the experiment were 21 students of Media Computer Sciences at Berlin University of Applied Sciences in their second year. They were informed that the experiment dealt with the quality of synthetic speech, but not about the details of parameters manipulated. Figure 2 displays a map of the stimuli created for the experiment. It becomes clear that not all possible combinations of durational and $F0$ modifications were tested. As the experiment was performed during a scheduled lecture, a maximum experiment duration of 90 minutes had to be observed.

Informal listening tests had shown little degradation for stimuli with extracted Fujisaki parameters ($rho_{F0} \approx 0.93$), as well as for those with durational correlations of $rho_{\text{dur}} = 0.9$. Therefore in these cases only three combinations were created, otherwise five.

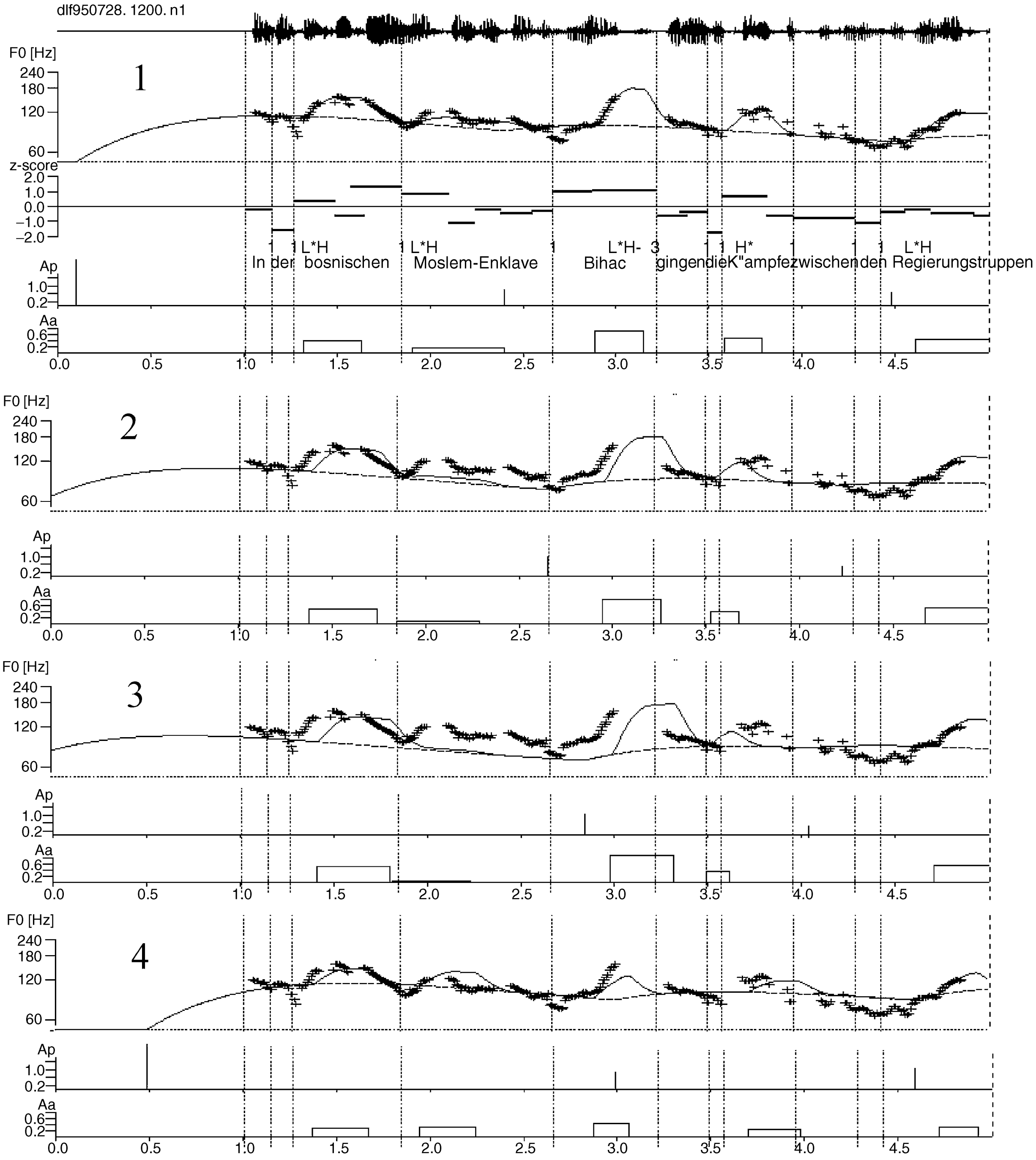

Figure 1. Four examples of $F0$ contours used in the perception experiment. The top panel (1) displays from top to bottom: the original speech waveform, the extracted (+) and model-generated $F0$ contours (solid line), duration contour (syllabic z-score), ToBI tier, text of utterance, underlying phrase and accent commands. Utterance displayed: "In der bosnischen Moslem-Enklave Bihac gingen die Kämpfe zwischen den Regierungstruppen…"-"In the Bosnian Muslim-enclave of Bihac, fighting between the government troops…". The bottom panels display $F0$ contours and underlying Fujisaki model commands for the following cases: (2) degraded Fujisaki parameters, $rho_{F0} = 0.7$; (3) degraded Fujisaki parameters, $rho_{F0} = 0.5$; (4) Fujisaki parameters predicted by the integrated model. It can be seen that the amplitudes of predicted accent commands (Panel 4) show less variation than the extracted ones (Panel 1).

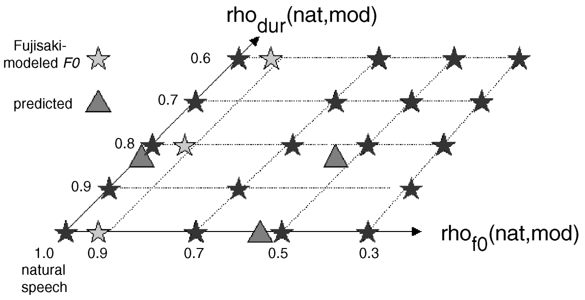

Figure 2. Array of stimuli used in the perception experiment. The dark-gray triangles denote versions using integrated model-based prediction, one with predicted durations and natural $F0$, one with natural durations and predicted $F0$, and a third one using both durations and $F0$ from the prediction. The light-gray stars denote stimuli produced using extracted Fujisaki-parameter configurations.

For the stimuli generated using the integrated model, an average rho_{F0} of 0.55 was calculated, and an average rho_{dur} of 0.82. As can be seen from Fig. 2, for each sentence, 25 different versions were created, yielding a total number of 300 stimuli. In order to test the consistency of the quality judgment, stimuli pertaining to four of the sentences were included twice, bringing the total number of stimuli to 400. 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'.

As a brief introduction to the 'quality spectrum' of the stimuli, and also in order to familiarize the subjects with the voice characteristics of the speaker, an original recording of a news story was played back (supposed to receive a judgment of '1'). In the following a random choice of 10 stimuli with manipulated prosodic parameters of varying quality was presented.

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. After presenting the first half of the stimuli, a five-minute break was taken.

The grading approach which obviously holds the risk of less accurate judgments compared with A/B comparisons as performed in Mixdorff and Mehnert (1999), was chosen because of the large number of stimuli involved, since A/B comparison between 25 different stimuli would imply 25! decisions times the number of sentences, and even A/B comparison between adjacent stimuli only would have implied a number of about 100 decisions per sentence.

With a total number of 400 stimuli presented in 90 minutes, and considering the numerous repetitions of the same sentence, the cognitive load on the subjects was extremely high. Especially the stimuli with extremely low correlation values of $rho_{\mathrm{dur}} = 0.6$ and $rho_{F0} = 0.3$ which had been introduced to the set of utterances in order to mark the 'very bad' end of the quality spectrum, sometimes raised amused reactions because of their strongly distorted prosody.

6.2. Results from the First Experiment

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* over all sentences (first presentations only) is displayed in Table 3 which shows that the *MOS* generally decreases with deteriorating prosody, with the steepest decay found between a rho_{dur} of 0.7 and 0.6.

The natural speech stimuli were not unanimously rated 'very good' and only reach a *MOS* of 4.37, with an average of 7.5 of the 16 'original' stimuli in the set of utterances being assigned 'grade 1'.

Furthermore it can be seen that the scores for the stimuli produced with the integrated model slightly exceed those assigned to adjacent 'degraded stimuli'. The *MOS* for the integrated model (predicted durations, natural $F0$) of 4.18 compares to that of $rho_{\mathrm{dur}} = 1.0/rho_{F0} = 0.9$ (4.13). In the case of integrated model-based $F0$ and natural durations, the score of 3.83 compares to $rho_{\mathrm{dur}} = 1.0/rho_{F0} = 0.7$ (3.82). Joint predictions of duration and $F0$ (3.40) is perceptually closest to the 'degraded stimuli' of $rho_{\mathrm{dur}} = 0.8/rho_{F0} = 0.9$ ($MOS = 3.33$).

Table 3. Overview of MOS results for the first experiment averaged over all 12 sentences and all subjects, first presentation.

rho_{f0} → rho_{dur} ↓	1.0	0.9	0.7	0.5	0.3	IGM
1.0	4.37	4.13	3.82	3.66	3.50	3.83
0.9	3.94	–	3.54	–	3.29	–
0.8	3.57	3.33	3.17	2.96	2.79	–
0.7	3.07	–	2.73	2.57	2.53	–
0.6	2.35	2.10	2.04	1.93	1.88	–
IGM	4.18	–	–	–	–	3.40

Table 4. Overview of MOS results averaged over all 12 sentences and all subjects, first presentation.

rho_{f0} → rho_{dur} ↓	1.0	0.9	0.7	0.5	IGM
1.0	4.66	4.50	3.97	3.70	3.58
0.9	4.40	4.02	3.80	3.51	–
0.8	3.70	3.36	2.94	2.89	–
0.7	2.70	2.63	2.30	2.17	–
IGM	4.13	–	–	–	2.99

Au: pls. cite table 4 in the text.

Factor analysis, excluding stimuli produced with the integrated model, shows a correlation between the *MOS* and rho_{dur} of 0.79, and a correlation between *MOS* and rho_{F0} of 0.29 ($p < 0.01$ for both factors), suggesting that duration is the predominant factor for the quality judgment. The identity of the sentence was identified as a secondary factor ($p < 0.05$). No significant correlation, however, was found between the *MOS* and the length of a sentence in terms of the number of syllables it contains.

Although all subjects actually made use of all available grades from 1 to 5, the individual averages vary between 2.14 and 3.85 (mean = 2.87, slightly less than the scale means of 3.0). The mean inter-subject correlation was found to be of 0.63, indicating considerable individual variations. The correlation between ratings at first and second time of presentation amounts to 0.923 indicating a relatively high intra-subject consistency.

6.3. *Sentences not Pertaining to the News Corpus*

The second experiment of the perceptual evaluation focussed on the performance of the integrated prosodic model on sentences which do not pertain to the news corpus. For this purpose, six sentences were chosen from the D set used in Experiment 1 and complemented by the following six sentences which had been used in Mixdorff and Mehnert (1999), henceforth called the 'J set'.

1. Bereitwillig gab er Auskunft.—*Willingly he supplied information.*
2. Aller Anfang ist schwer.—*Starting-off is always difficult.*
3. Die Begründung ist stichhaltig.—*The reason is up to the point.*
4. Das Gespräch zeigte die Gegensätze und die gemeinsamen Züge unserer Auffassungen.—*The conversation showed the opposites and common grounds of our views.*
5. Wenn wir die Maschine anschliessen, beginnt der Motor zu surren.—*When we connect the machine, the engine will start humming.*
6. Es regnete soviel, dass der Fluss über die Ufer trat.—*It rained so much that the river burst its banks.*

The setting was similar to that of the first experiment. Stimuli produced with IGM were presented in a reference matrix of degradation stimuli. As for the versions created with the IGM, variants with predicted durations and $F0$ contour, as well as variants with either of the features copied from the natural utterances were produced. Degradation stimuli were created with a minimum of rho_{dur} of 0.7, and a minimum of rho_{F0} of 0.5.

These values were chosen as the stimuli with lower correlation coefficients used in Experiment 1 had sounded overly distorted. Figure 3 displays the total matrix of stimuli employed.

In order to produce the degraded stimuli, the original recordings from the J set were analyzed using the automatic approach for Fujisaki parameter extraction. For the speaker of the J set, an *Fb* of 75 Hz and an α of 2/s was determined. Creating the IGM-based stimuli for the J set posed certain problems: As the α of the news speaker was only 0.95/s, and *Fb* as low as 50.2 Hz, the phrase command amplitudes *Ap* output by the IGM were considerably higher than those required for an α of 2/s. This relationship results from the formulation of the phrase control mechanism: It is proportional to the

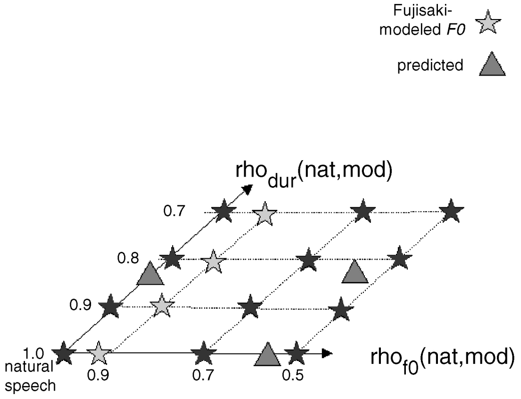

Figure 3. Array of stimuli used in Experiment 2. The dark-gray triangles denote versions using integrated model-based prediction, one with predicted durations and natural $F0$, one with natural durations and predicted $F0$, and a third one using both durations and $F0$ from the prediction. The light-gray stars denote stimuli produced using extracted Fujisaki parameters.

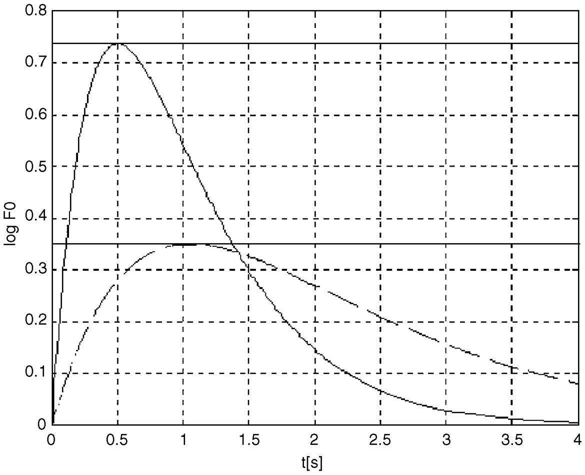

Figure 4. Phrase component for α of 0.95 (dashed) and 2.0 (solid). The maximum values of $F0$ (marked by horizontal lines), which roughly coincide with the segmental onset of a phrase are 0.35 and 0.74, respectively.

square of α. Hence, for a given maximum $F0$ modification caused by the phrase component (i.e. the log $F0$ reached on the top of the phrase component, see Fig. 4, a higher α requires a lower *Ap* and vice versa. Preserving an α of 2/s and re-scaling the phrase command amplitudes *Ap* with a constant factor did not solve the problem, as consecutive commands (due to the steeper slope at a higher α) were too low to compensate for the declination effect. However, when *Fb* and α were both taken from the news speaker D, the resulting $F0$ contours were too low for speaker J.

Eventually, a compromise was found by keeping an α of 0.95/s and scaling *Ap* to match an *Fb* of 75 Hz. The scaling factor was empirically determined by adjusting the mean $F0$ of the model-generated contours to that of the extracted natural contours.

Syllable durations for the J set were calculated using the IGM by supplying the underlying linguistic and phonetic information, i.e. the 24 input parameters of the IGM. Calculation showed that the correlation between measured and predicted syllable durations on the J set was of 0.86.

Subjects taking part in the experiments were 20 members of staff and students of the Laboratory of Acoustics and Speech Communication, Dresden University of Technology, 16 males and four females. All of them were experienced in listening to synthetic speech. For logistic reasons, unlike Experiment 1, Experiment 2 was performed on a PC using headphones.

After being questioned their names, subjects were presented the randomized stimuli. Each stimulus was played back twice, then the subjects had to specify a grade between 1 and 5 (very good . . . very bad).

The choices were automatically logged to a protocol file. The sequence of the first 30 stimuli was presented twice, in order to familiarize the subjects with the quality spectrum of the stimuli.

6.4. *Results from the Second Experiment*

Table 5 lists the mean opinion scores averaged over all subjects and the total of twelve sentences for the different stimulus conditions. As in the first experiment, the MOS was determined by subtracting the grades assigned by the ubjects from 6, yielding a scale ascending with quality between 1 and 5.

As can be seen from the table, the *MOS* monotonously decreases with rho_{dur} and rho_{F0}, but the effect is much stronger in the direction of rho_{dur}. This is also reflected by the correlation coefficients between *MOS* and rho_{dur} and rho_{F0}, which are of 0.78 and 0.36, respectively.

Correlation analysis shows that the *MOS* is strongly influenced by the corpus ($\rho = 0.19$, $p < 0.07$ for the degraded stimuli, and $\rho = 0.53$, $p < 0.01$ for the model-generated parameters). If we display the results for D set and J set separately (see Table 6), it becomes clear that the samples from the J set were generally assessed to be of poorer quality. This observation not only concerns stimuli produced with the IGM, but even the originals. The quality distance, however, is greatest when the $F0$ contours are supplied by the IGM. In the case of syllable durations and $F0$ contours supplied by the integrated model for the J Set, it is judged more poorly (*MOS* = 2.57) than reference stimuli with rho_{dur} of 0.8 and rho_{F0} of 0.5 (2.71). If syllable durations only are taken from the IGM (*MOS* = 3.81),

Table 5. Overview of MOS results averaged over all 12 sentences and all subjects.

rho_{F0} → rho_{dur} ↓	1.0	0.9	0.7	0.5	IGM
1.0	4.66	4.50	3.97	3.70	3.58
0.9	4.40	4.02	3.80	3.51	–
0.8	3.70	3.36	2.94	2.89	–
0.7	2.70	2.63	2.30	2.17	–
IGM	4.13	–	–	–	2.99

Table 6. Overview of MOS results for J set and D set listed separately.

$rho_{f0} \rightarrow$ $rho_{\mathrm{dur}} \downarrow$	1.0	0.9	0.7	0.5	IGM
J set					
1.0	4.55	4.26	3.69	3.40	3.06
0.9	4.33	3.80	3.55	3.18	–
0.8	3.61	3.23	2.85	2.71	–
0.7	2.56	2.47	2.27	2.06	–
IGM	3.81	–	–	–	2.57
D set					
1.0	4.77	4.73	4.26	4.00	4.11
0.9	4.46	4.24	4.05	3.85	–
0.8	3.79	3.49	3.02	3.07	–
0.7	2.84	2.80	2.32	2.28	–
IGM	4.44	–	–	–	3.40

however, it is rated better than degraded stimuli at a rho_{dur} of 0.8 (3.61). IGM-predicted $F0$ contour values only ($MOS = 3.06$) are rated even lower than the corresponding reference stimuli for a rho_{F0} of 0.5. This confirms that the $F0$ contours for the J set were judged inadequate by the listeners. Furthermore, the results for the D set better correspond to those obtained in the first experiment.

The intra-speaker correlation on the 30 stimuli that were presented twice was of 0.94, and the inter-speaker correlation of 0.64, which is similar to that found in the first experiment.

7. Discussion and Conclusions

The current study introduced and examined an 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 MFN predicting syllable durations and energy along with syllable-aligned Fujisaki control parameters.

A method for evaluating the approach within a matrix of stimuli produced by controlled degrading of the natural utterances was tested. Results indicate that subjects are more sensitive to errors in the prediction of syllable durations than to variations in the $F0$ contour, though errors in the duration domain will also negatively affect the $F0$ contours.

In terms of predicted syllable durations, the integrated model comes close to natural durations whereas the prediction of Fujisaki control parameters, especially *Aa*, obviously requires additional input parameters. This becomes evident with the results of statistical analysis discussed in Section 4.

The results yielded in the second experiment shows that the IGM, at least as far as the prediction of $F0$ is concerned, performs more poorly on a speaker with different *Fb* and time constant α than the one on which it was trained, probably mostly due to the compatibility problem which cannot be satisfactorily solved by simply scaling the phrase command magnitude *Ap*.

A further reason for the lower ratings of the J set might be the fact that the speaker of the J set himself was a member of staff of the Dresden laboratory for many years, and produced speech corpora and an inventory for diphone synthesis. Hence his voice quality and register were extremely familiar to all of the subjects in Experiment 2. The assumption that this ‘conditioning’ influenced the listeners’ judgment is supported by results reported in Mixdorff and Jokisch (2001b), where stimuli from the D and J sets yielded similar results when judged by subjects not familiar with speaker J, that is, no significant correlation was found between the *MOS* and the corpus. Considering these observations, using known and unknown speakers in the same perception test should be avoided.

The method of controlled degradation appears to be a useful paradigm to be applied to the evaluation of synthetic prosody as mean opinion scores can to a certain extent be predicted from the correlation in the duration and $F0$ domains.

A further advantage is that the segmental quality, which considerably deteriorates when diphone synthesis is used, remains relatively high under resynthesis conditions with less bias on subjects’ judgments on prosody. Compared with restricted representations (see, for instance, Sonntag and Portele, 1998), prosodic degrading has the advantage of presenting subjects with real speech stimuli and not ‘quasi-speech signals’ not conveying any meaning and void of segmental information. Still the number of stimuli presented in one session should be smaller than that in the first experiment, as can be seen from the relatively small number of original stimuli which were assigned ‘grade 1’.

Further evaluation of results from the current perception experiments is needed in order to explain scores for individual stimuli that diverge from the norm and identify shortcomings of the integrated approach.

Future work will concern the introduction of further potential predictive factors, such as the syntactic environment and the focal condition, and the integration of the prosodic module into the TTS system DreSS.

Note

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.

References

Corrigan, G., Massey, N., and Karaaly, O. (1997). Generating segment durations in a text-to-speech system: A hybrid rule-based/neural network approach. *Proc. Eurospeech'97*. Rhodes, vol. 5, pp. 2675–2678.

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.

Hecht-Nielsen, R. (1990). *Neurocomputing*. Reading (Mass.): Addison-Wesley.

Hoffmann, R. (1999/II). A multilingual text-to-speech system. *The Phonetician*, *80*:5–10.

Jokisch, O., Hirschfeld, D., Eichner, M., and Hoffmann, R. (1998). Multi-level rhythm control for speech synthesis using hybrid data driven and rule-based approaches. *Proceedings of ICSLP'98*. Sydney, pp. 607–610.

Jokisch, O., Mixdorff, H. et al. (2000). Learning the parameters of quantitative prosody models. *Proceedings ICSLP 2000*. Beijing, China, vol. 1, pp. 645–648.

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

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

Mixdorff, H. and Jokisch, O. (2001a). Building an integrated prosodic model of German. *Proceedings of Eurospeech 2001*. Aalborg, Denmark, vol. 2, pp. 947–950.

Mixdorff, H. and Jokisch, O. (2001b). Comparing a data-driven and a rule-based approach to predicting prosodic features of German. *Tagungsband der 12. Konferenz Elektronische Sprachsignalverarbeitung*. Bonn, Germany, pp. 298–305.

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

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

Sonntag, G.P. and Portele, T. (1998). PURR—a method for prosody evaluation and investigation. *Journal of Computer Speech and Language*, *12*(4): 437–451. Special issue on evaluation in language and speech technology.

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

Zell, A. (1994). *Simulation Neuronaler Netze*, Bonn, Addison-Wesley.