Implementing and Evaluating an Integrated Approach to Modeling German Prosody

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 statistical results concerning the relationship between linguistic and phonetic information underlying an utterance and its prosodic features. These results were employed for training an FFNN-based integrated prosodic model predicting syllable duration and energy along with syllable-aligned Fujisaki control parameters. A novel method of perceptual evaluation was applied, 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 is a by far stronger factor with respect to the perceived quality than the accuracy of the predicted *F0* contour.

1. Introduction

It is widely acknowledged that the intellegibility 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 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 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.

2. Properties of the Integrated Model

The model is henceforth to be called an 'integrated prosodic model' [6], 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 data base.

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 *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 [7].

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 parameters alpha, beta and Fb are assumed to be constant for the same speaker.

Output Parameter of Model	Calculated as	N of tokens in data base
syllable duration	t_{off} - t_{on}	13151
Aa	-	3022
$T1_{dist}$	$T1$-t_{on}	3022
$T2_{dist}$	$T2$-t_{off}	3022
Ap	-	1047
$T0_{dist}$	t_{on}-$T0$	1047
energy	mean frame power *rms* in syllable	13151
pause	inter-phrase pause duration	1047

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 [8], 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 [9]. 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

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), stressed, but unaccented (1), or stressed and accented (2), i.e. bearing a tone switch (ρ=.349). As expected, 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).

Relationships established between linguistic/phonetic factors and Fujisaki control parameter are generally in line with the results of earlier works ([3], 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 (ρ=.257) whereas the part-of-speech of the superordinate word has relatively little influence (ρ=.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 data base, as well as a more detailed description of the syntactic environment.

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, 20 syllabic features were selected as 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 ($T0_{dist}$, *Ap*). The output vector consists of five Fujisaki model

[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.

parameters with relative timing controlling the *F0 contour* (*T1*$_{dis}$, *T2*$_{dis}$, *Aa*, *T0*$_{dist}$, *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 fully-connected feed-forward neural network (FFNN) of four 4 layers (24x18x12x8 neurons). 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 (Stuttgart corpus) and standard error backpropagation, minimizing the root mean square error (RMSE) between teaching input and net output.

The Stuttgart corpus was subdivided into a training set (10.000 syllables) and an independent test set (3151 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 data base from which they are computed (see Section 4). The following table shows the RMSE, means and standard deviations of trained and predicted output parameters.

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

Param-eter	**overall RMSE**	**mean (trained)**	**s.d. (trained)**	**mean (pred.)**	**s.d. (pred.)**
T1$_{dist}$	0.058 s	.037 s	.152 s	.051 s	.097 s
T2$_{dist}$	0.068 s	.039 s	.181 s	.052 s	.118 s
Aa	0.059	.293	.165	.273	.064
T0$_{dist}$	0.138 s	.435 s	.544 s	.432 s	.196 s
Ap	0.140	1.10	.62	1.08	.57
syllab.dur.	0.046 s	.189 s	.078 s	.191 s	.064 s
pause	0.069 s	.290 s	.348 s	.290	.255 s
energy	689.8	1697.0	774.5	1677.2	360.4

It becomes clear that the predicted parameters exhibit reduced standard deviations, indicating the averaging behavior of the neural network.

After re-scaling all parameters, and relating the relative timing parameters of the Fujisaki model to the timing of the underlying syllable chain, the model commands can be fed into the Fujisaki model, producing a time-aligned *F0* contour.

The combination of a data driven approach, such as a neural network, with a rule-based prosodic model can be considered as a hybrid architecture (HYDRA, see also [6] which addresses the rapid adaptation to prosodic model parameters using a well-defined rule-based core).

6. Evaluating the Model

6.1. Experiment Design

A perception experiment for evaluating the quality of the prosodic model was designed. 12 sentences from the data base of varying complexity (between 13 and 44 syllables) 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 modelled 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.

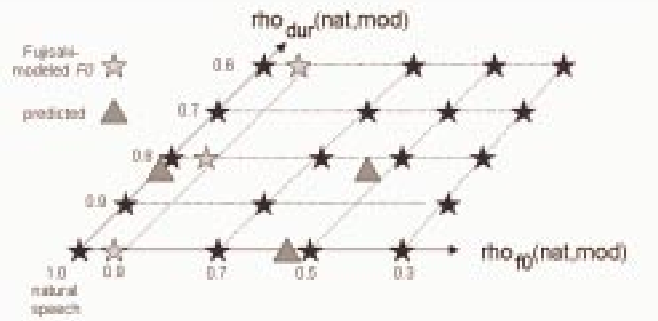

Figure 1:Array of stimuli used in the perception experiment. The dark-grey triangles denote versions using integrated model-based prediction, one with predicted durations and natural F0, one with natural durations and predicted F0, and a third one using both durations and F0 from the prediction. The light-grey stars denote stimuli produced using extracted Fujisaki-parameter configurations.

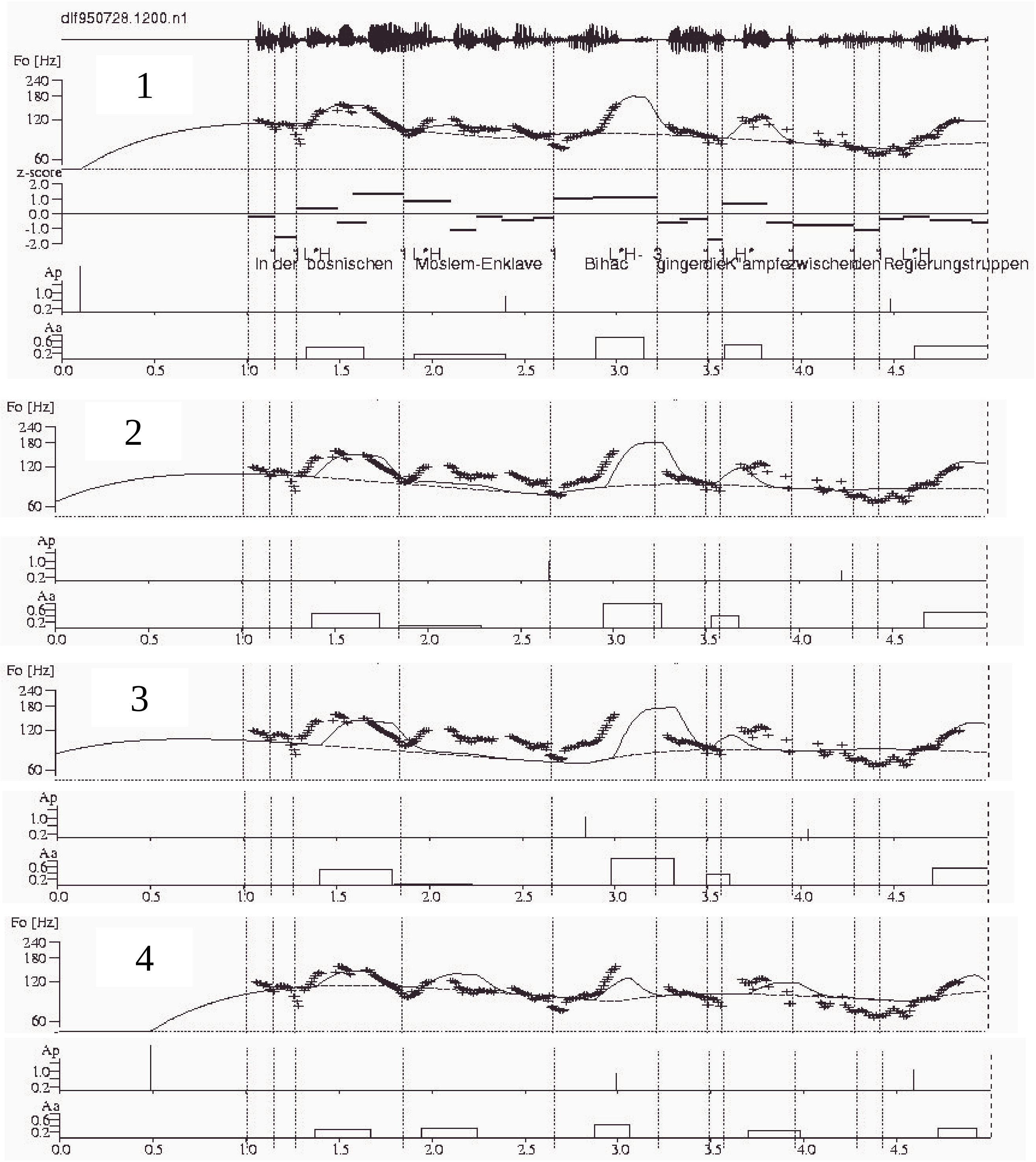

Figure 2: Four examples of F0 contours used in the perception experiment. The top panel (1) displays from top to bottom: the original speech waveform, the extracted (+) and model-generated F0 contours (solid line), duration contour (syllabic z-score), ToBI tier, text of utterance, underlying phrase and 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, $\boldsymbol{rho_{F0}}$ *= 0.7; (3) degraded Fujisaki parameters,* $\boldsymbol{rho_{F0}}$ *=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).*

In order to equally distribute the error onto all control parameters involved ($T1_{dis}$, $T2_{dis}$, Aa, $T0_{dist}$, 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.

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 1 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_{dur} = 0.9. Therefore in these cases only three combinations were created, otherwise five.

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 Figure 1, for each sentence, 25 different versions were created, yielding a total number of 300 stimuli. In order to test the consistency of the quality judgement, 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 [5], 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 students was extremely high. Especially the stimuli with extremely low correlation values of rho_{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. First Experiment Results

In order to yield a score increasing with perceived stimulus quality (*MOS*), grades assigned by the subjects were subtracted from a value of 5, producing a scale from 0 to 4. The *MOS* over all sentences is displayed in Figure 3, expressed by the height of arrows assigned to each stimulus in the array. The general tendency, that the *MOS* decreases with deteriorating prosody becomes obvious, 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 3.296, 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 exceed those assigned to adjacent 'degraded stimuli'. The *MOS* for the integrated model (predicted durations, natural *F0*) of 3.08 compares to that of rho_{dur} = 0.9 / rho_{F0}=1.0 (3.07). In the case of integrated model-based *F0* and natural durations, the score of 2.77 compares to rho_{dur} = 1.0/rho_{F0}=0.7 (2.78). Joint predictions of duration and *F0*, however, yield poorer results that the corresponding 'degraded stimuli': 2.38 against 2.50 for rho_{dur} = 0.9/rho_{F0}=0.7.

Factor analysis, excluding stimuli produced with the integrated model, showed a correlation between the *MOS* and rho_{dur} of 0.79, and a a correlation between *MOS* and rho_{F0} of 0.29 ($p < 0.01$ for both factors), indicating 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 0.632, indicating considerable individual variations.

The correlation between ratings at first and second time of presentation amounts to 0.923 indicating a high intra-subject consistency.

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 FFNN predicting syllable durations and energy along with syllable-aligned Fujisaki control parameters.

A novel method for evaluating the approach within a matrix of stimuli produced by controlled degrading of the natural utterances was tested. First results indicate that subjects are far more sensitive to errors in the prediction of syllable

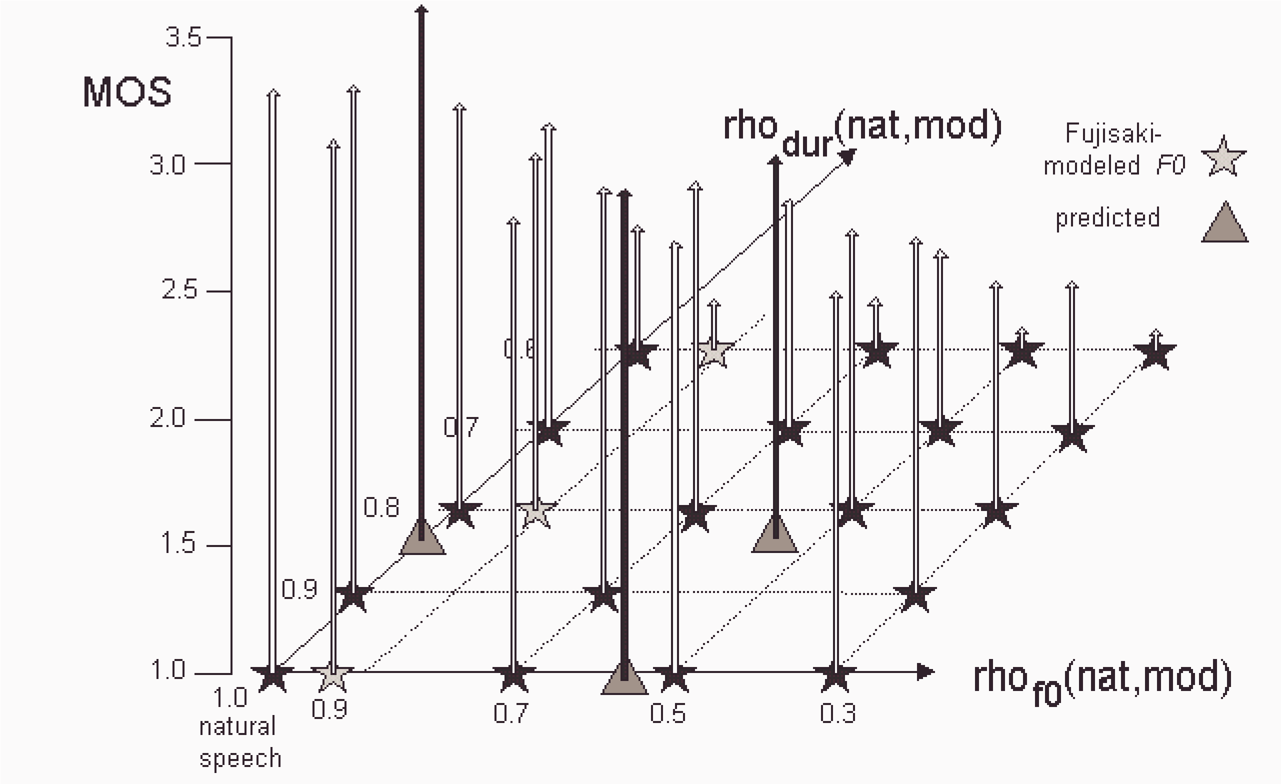

Figure 3: Total result of the perception experiment. For each stimulus in the array the corresponding mean opinion score (MOS) is indicated by the height of the vertical arrow. It becomes clear that model-based predictions generally yield higher ratings than stimuli created by controlled degrading with comparable cross-correlation values.

durations than to variations in the *F0* contour. 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 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.

Further evaluation of results from the currrnt perception experiment is needed in order to explain scores for individual stimuli and identify shortcomings of the integrated approach. Future work will concern 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.

8. 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] Mixdorff, H. and Fujisaki, H., "A quantitative description of German prosody offering symbolic labels as a by-product", in *Proceedings ICSLP 2000,* vol. 2., 98-101. Beijing, China, 2000.

[7] 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.

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

[9] 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.