LEARNING THE PARAMETERS OF QUANTITATIVE PROSODY MODELS

Oliver Jokisch, Hansjörg Mixdorff, Hans Kruschke, Ulrich Kordon

Laboratory of Acoustics and Speech Communication, Dresden University of Technology
D-01062 Dresden, Germany
Email: jokisch@eakss2.et.tu-dresden.de

ABSTRACT

The article introduces a novel hybrid data driven and rule based approach for the prosody control in a TTS system, which combines the advantages of well-balanced, quantitative models with the flexible training of derived model parameters. Instancing the training of Fujisaki intonation parameters for German (MFGI) the article describes the hybrid data driven and rule based architecture HYDRA, the speech database, the extraction of the model parameters and the neural network (NN) training of these parameters. Preliminary results using the hybrid intonation model are presented. A hybrid neural network and rule based, quantitative model can be easily parameterized and adapted e.g. for multilingual applications, but has a higher complexity and requires the automatic extraction of the model parameters from a speech database.

1. INTRODUCTION

The synthesis of near-to-natural prosodic contours is still an important issue in text-to-speech (TTS). Several studies, such as [2] prove the strong effect of the synthetic prosody on naturalness and intelligibility of synthetic speech. Focusing on the F0 contour and segmental durations the prosody structure of synthetic speech signals can be parameterized by established quantitative models. Data driven algorithms for prosody control enable the simple adjustment of prosodic parameters via training and the generation of more variable contours. Nevertheless, a strictly data driven approach using e.g. a neural network (NN) as in [3] tends to local runaways and similar irregularities. This contribution introduces a hybrid neural network and rule based approach, which combines the advantages of well-balanced, quantitative models with the flexible training of derived model parameters. The article starts with an illustration of the hybrid data driven and rule based prosody model and of the TTS target system (chapter 2). The 3rd chapter deals with the speech corpora used for statistical analyses and to extract duration and intonation parameters. The 4th chapter describes a feed forward NN for predicting MFGI parameters. Finally, preliminary results of the hybrid intonation model are presented.

2. THE HYBRID DATA DRIVEN AND RULE BASED PROSODY MODEL

In past a number of established rule based and some data driven prosody models were developed. Basing on extensive research work - rule based, quantitative prosody models often outperform strictly data driven approaches [4] (Listeners preferred Fujisaki versus NN generated intonation contours.). Keeping the features of rule based modelling it seems to be necessary to extend the models by parameter-learning and "adjusting components". In [5] a hybrid data driven and rule based approach for the duration control in the TTS system DRESS has been presented.

2.1. Approach

The hybrid data driven and rule based architecture (HYDRA) includes a rule based, quantitative prosody model (the core component), an automatic extractor for the model parameters, a data driven algorithm for learning and adjusting the model parameters and an interface to the according databases (training time) respectively to the TTS text processor at work (figure 1).

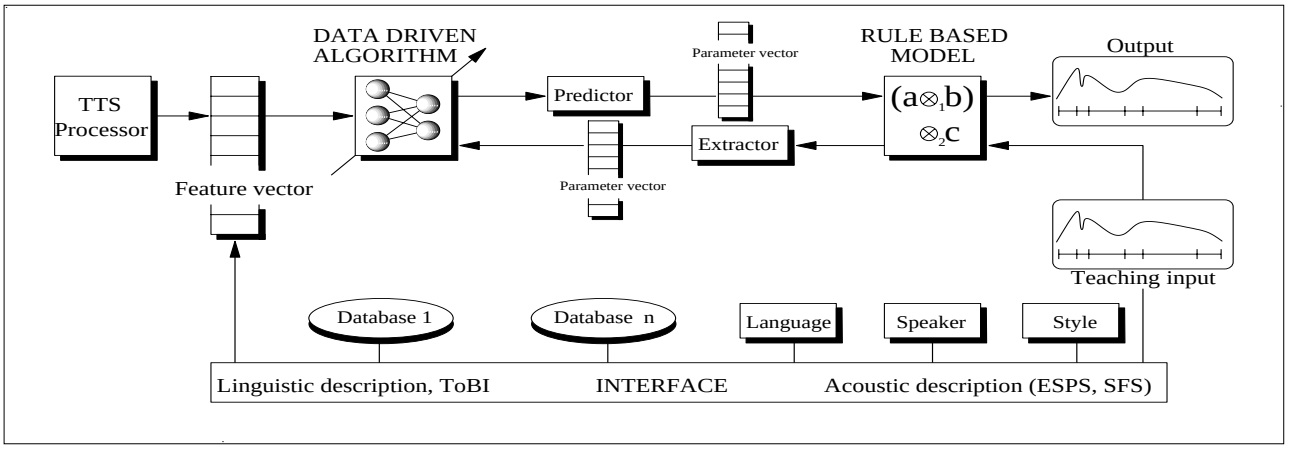

Figure 1. Hybrid data driven and rule based architecture (HYDRA) for generating the prosody in a TTS system

During the training a linguistic-phonetic feature vector is applied to the input of the data driven module while the output is predicting the parameter vector for the input of the rule based module (supervised learning). The parameter vector may consist of initial times, amplitudes, intensities or similar signal parameters.

In opposite to other data driven approaches for the prosody control focusing on "a learning procedure including an output model" the HYDRA approach is designed as "an output model extended by a learning procedure". The approach underlines the importance of a well-balanced rule based model and limits the variety of possible outputs. The data driven module acts as a "controller" and enables the adjustment of mean values, amplitudes, etc. for new languages, speakers or speaking styles. The model parameter extractor itself may use a learning procedure, as well.

2.2. Application in the Text-to-Speech System DRESS

The hybrid data driven and rule based approach described for generating the prosody is implemented in the Dresden TTS system DRESS (Dresden Speech Synthesizer).

DRESS is a time-domain synthesizer (multiphone and syllable units) with a preprocessing module, diverse lexicons, a grapheme-phoneme converter, replaceable modules for duration and intonation control and an acoustic module. Recent improvements refer to multilinguality and a better naturalness. Multilingual databases for German, US-English, Russian, Italian, Czech and Chinese have been collected and connected to DRESS.

Processing a rule-based generated phoneme sequence - enriched with tagged and classified accents, syllables, words, phrases and sentences – DRESS offers 3 alternative duration models (rule based, NN [6], m-level [5]) and 3 intonation models (linear, NN [6], MFGI [1]).

The HYDRA concept applied to the MFGI model results in a novel combined NN-MFGI intonation model described in 4.1.

3. SPEECH DATABASE

The database used in this study is part of a German speech corpus compiled by the Institute of Natural Language Processing at the University of Stuttgart [7]. It consists of 72 broadcasting news stories read by a male speaker. The total recording time includes 48 minutes of speech containing 13151 syllables. The corpus contains boundary labels for phones, syllables and words as well as ToBI-labels following the Stuttgart System [8].

Additionally, a Dresden database containing fairy tales (5043 syllables, male and female speaker) will be extended according to the Stuttgart conventions to supply validation data for the created hybrid models and to allows the training of different speaking styles.

4. NN-PREDICTOR FOR THE MFGI-FUJISAKI PARAMETERS

Following the HYDRA approach a NN predictor for the Mixdorff-Fujisaki German Intonation (MFGI [1]) model parameters can be created.

The necessary, initial extractor of the MFGI parameters is explained in [9]: A given F0 contour is approximated by quadratic splines. The resulting, stylized F0 contour is high-pass filtered (HF contour) and subtracted from the spline contour (LF contour). The overall minimum of the LF contour is initially set to Fb. By searching local extreme values the parameters T0, Ap, T1, T2 and Aa can be initialized. Iterative, the overall mean-square error is minimized using the Analysis-by-Synthesis procedure. The extractor produces reliable results (See also table 1 in chapter 5).

4.1 Method

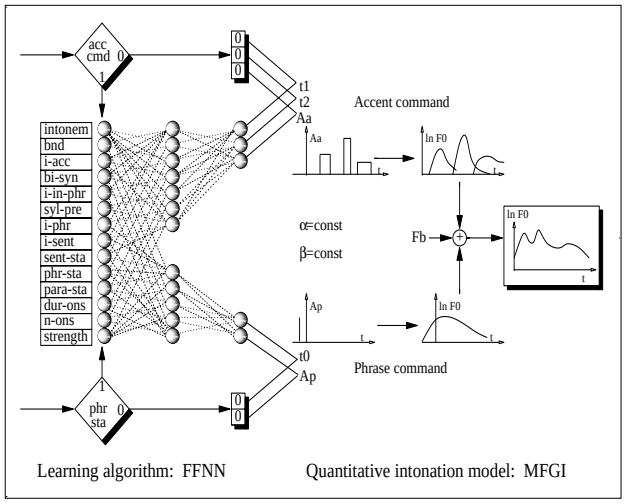

Figure 2. FFNN-MFGI-Predictor: Estimating the intonation model parameters for each input syllable

The MFGI parameters previously extracted are the teaching input for two fully-connected feed-forward NN (FFNN). The first FFNN is predicting the accent command (T1,T2, Aa) - the second one the possible phrase command (T0, Ap) - for the respective syllable (figure 2). In case of an active accent flag or a phrase-start flag a vector of 14 features describing the prosodic context of the syllable (phrase and sentence position, accent position and strength, onset duration, etc.) is applied to the FFNN input layer. The syllable sequence of the output vectors (T1, T2, Aa) and (T0, Ap) provides the parameter set for the MFGI model. Considering a single phrase α, β and Fb are given by an additional input channel (mean value assumption). According to the Fujisaki approach [1] the resulting, separate accent and phrase commands are super-positioned.

4.2 NN-Training

The Stuttgart corpus was subdivided into three sets containing training sample (5000 syllables), test sample (5000 syllables) and validation sample (3151 syllables). Considering the mentioned flags for accent or phrase commands the final learning and testing sets are only subsets of those samples (e.g. accent command patterns: 1165, phrase command patterns: 522 in the training sample). The patterns are trained using Error-Backpropagation and minimizing the Root Mean Square Error (RMSE) between the teaching sequence of parameters (using the MFGI extractor) and the FFNN predicted parameter sequence.

The FFNN training is carried out using the training sample. The test sample's RMSE defines the stop criterion (avoiding over-adaptation) and the validation sample enables an independent evaluation.

5. PRELIMINARY RESULTS

For a first evaluation of the introduced NN-MFGI predictor for the German intonation a few observations and measurements are presented in the following chapter. Figure 3 shows the differences between the extracted ("parameterized") and the NN predicted phrase respectively accent commands according to the MFGI model (typical example). The NN seems to learn the basic concept of the MFGI model but partially produces higher differences to the teaching input (e.g. timing and amplitudes of the 2nd phrase command). Considering the ambiguity of the Fujisaki model (different parameter sets may produce a similar F0 output contour) the NN-MFGI predictor, nevertheless, generates proper overall results.

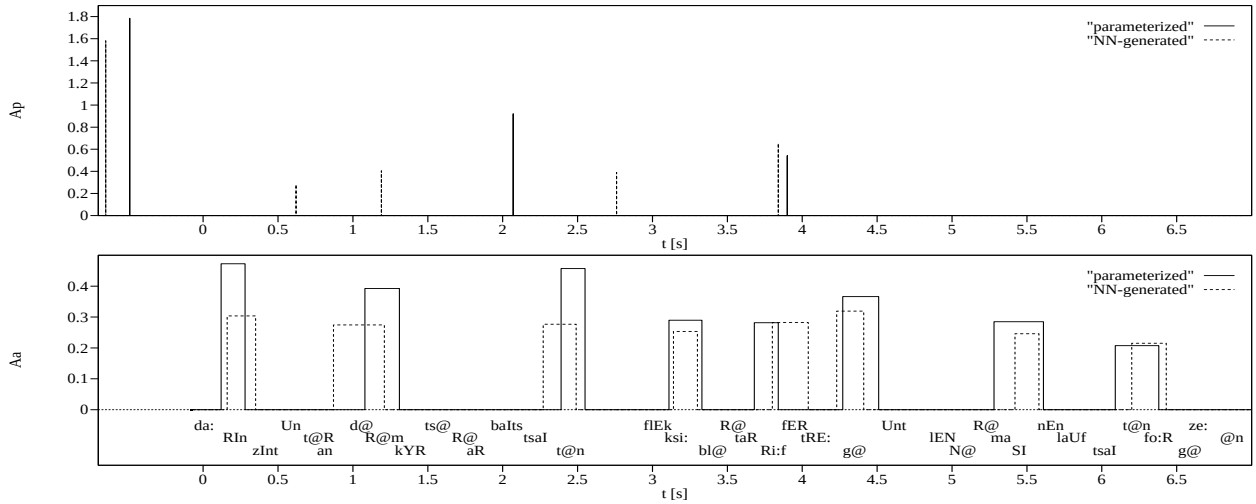

Figure 3. Extracted vs. predicted MFGI-model parameters (test sample). Top: phrase commands. Bottom: accent commands. Text: "Darin sind unter anderem kürzere Arbeitszeiten, flexiblere Tarifverträge und längere Maschinenlaufzeiten vorgesehen."

Table 1 summarizes the RMSE observed between the measured ("original") F0 contour, the MFGI-extracted and reconstructed contour and the NN-MFGI predicted contour. Considering the same sample (e.g. the test sample) the extracted and reconstructed contour obviously has a smaller RMSE of 14.2 Hz than the NN predicted contour (RMSE=17.8 Hz). May a student better perform than his teacher? In fact, the NN predictor matches the original contour better than its teaching input (RMSE=17.8 Hz vs. RMSE=19.2 Hz). Probably, the NN detects some contradictions in the extractor method or in the database and generalizes in such cases.

F0 Contour Set	RMSE [Hz] (Training)	RMSE [Hz] (Test)	RMSE [Hz] (Validation)
MFGI vs. original	16.7	14.2	12.8
NN vs. original	20.3	17.8	18.6
NN vs. MFGI	21.3	19.2	20.7

Table 1. RMSE of different F0 contour sets

Figures 4 and 5 present typical constellations of the test sample. The NN predicted F0 contour fits the global shape given by the original respectively the MFGI-reconstructed ("parameterized") contour fairly well. Local differences occur for stronger accents. The results for the training set and the validation set are similar. In particular, for learning the correct accent timing more input features are necessary.

The re-synthesis with the NN predicted contours versus the MFGI generated shows the applicability of the proposed method.

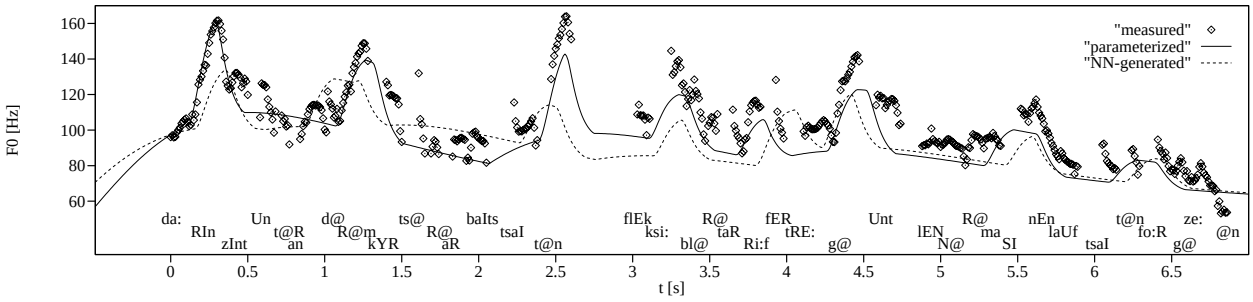

Figure 4. F0 contours basing on measured, extracted MFGI parameters and NN predicted MFGI parameters. Utterance: "Darin sind unter anderem kürzere Arbeitszeiten, flexiblere Tarifverträge und längere Maschinenlaufzeiten vorgesehen."

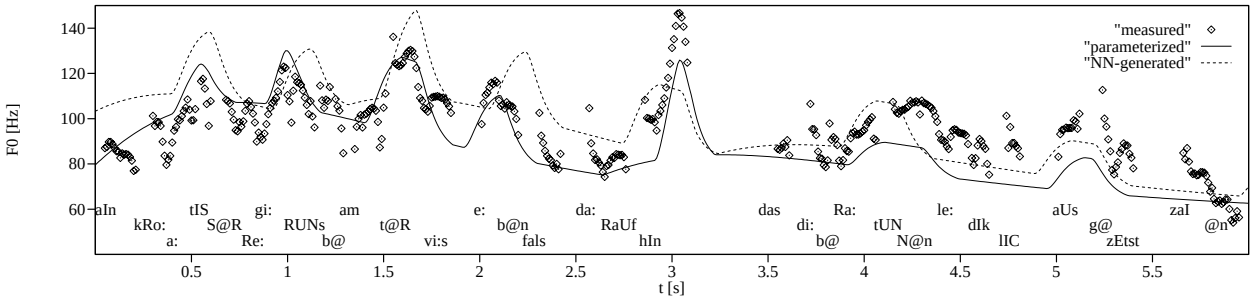

Figure 5. F0 contours basing on measured, extracted MFGI parameters and NN-predicted MFGI parameters. Utterance: "Ein kroatischer Regierungsbeamter wies ebenfalls darauf hin, daß die Beratungen lediglich ausgesetzt seien."

6. CONCLUSION

The proposed hybrid data driven and rule based approach (HYDRA) and its example for the NN prediction of MFGI parameters is a practicable solution to combine a good model performance with self-learning components. Further studies are required to improve the components of HYDRA, e.g.:

- the parameter extraction method
- the learning algorithm
- the database and the linguistic-phonetic feature vector

7. REFERENCES

1. Mixdorff, H.: Intonation patterns of German – Quantitative analysis and synthesis of F0 contours. Ph.D. thesis, TU Dresden, Dresden, 1998.
2. G. Sonntag et al: Comparative evaluation of six German TTS systems. *Proc. Eurospeech'99*, Budapest, Vol. 1, 251-254.
3. C. Traber: F0 generation with a database of natural f0 patterns and with a neural network. In G. Bailly, ed.: Talking Machines: Theories, Models and Designs, 287-304, North Holland, 1992.
4. R. Hoffmann et al: Evaluation of a multilingual TTS system with respect to the prosodic quality. *Proc. ICPhS'99*, San Francisco, 2307-2310
5. O. Jokisch et al: Multi-level rhythm control for speech synthesis using hybrid data driven and rule-based approaches. *Proc. ICSLP'98*, Sydney, 607-610.
6. O. Jokisch et al: Neuronale Prosodiegenerierung - Einfluss der Trainingsdaten. *Proc. DAGA'98* (24th German Conference on Acoustics), Zurich, 352-353.
7. Rapp, S.: Automatisierte Erstellung von Korpora für die Prosodieforschung. Ph.D. thesis, University of Stuttgart, Stuttgart, 1998.
8. Mayer, J.: Transcription of German intonation: The Stuttgart system. Technical report, University of Stuttgart; Stuttgart, 1995.
9. Mixdorff, H.: A novel approach to the to fully automatic extraction of Fujisaki model parameters. *Proc. ICASSP 2000*, Istanbul, Vol. 3, 1281-1284.