Evolutionary Optimization of an Adaptive Prosody Model

Oliver Jokisch and Michael Hofmann

Laboratory of Acoustics and Speech Communication
Dresden University of Technology, D-01062 Dresden, Germany

oliver.jokisch@ias.et.tu-dresden.de

Abstract

The perceived quality of synthetic speech strongly depends on its prosodic naturalness. Concerning the control of duration and fundamental frequency in a speech synthesis system, sophisticated models have been developed during the last decade. Departing from the syllable-based, adaptive prosody model IGM the authors surveyed a novel evolutionary approach to optimize the model structure itself and to finally improve the predicted prosodic contours.
Therefore, a German newsreader corpus has been trained using a feed forward neural network. In parallel, network and data configurations were automatically optimized using the Strength Pareto Evolutionary Algorithm (SPEA).
Achieving similar prediction results as in the original IGM configuration, the evolutionary optimization reduces the network complexity, in particular, the number of necessary input parameters from 24 to less than 10 by eliminating redundancies. This optimization method may be helpful in the further development of resource-saving prosody modules, e.g., for use in embedded text-to-speech applications and it also eases the difficult introspection of prosodic rules which are automatically generated during the training.
Nevertheless, preliminary perceptive tests show no significant differences in comparison to synthetic stimuli based on prosodic contours predicted by the original model.

1. Introduction

The Intelligibility and the perceived naturalness of synthetic speech strongly depend on the prosodic quality. Recent systems concatenating larger chunks of speech from a database achieve a considerably high quality, as they widely preserve the natural prosodic structure. Because of resource limitations prediction and control of prosodic parameters will keep their importance.
Adaptive (neural network) models for controlling segment durations or fundamental frequency were already suggested in 1992 from Campbell [1] and Traber [2]. Meanwhile, adaptive prosody models using a neural network core are widely accepted since they can be efficiently used in multilingual systems and also easily be adapted to realize new speaking styles or other individual characteristics.
Recently, Mixdorff and Jokisch suggested an integrated, adaptive model predicting f0, duration and intensity of syllables in German called IGM [3]. The neural network core of this model was successfully implemented into the Dresden Speech Synthesizer [4] and an analog design was tested for another language, as well [5].
There is a legitimate criticism concerning the difficult introspection of prosodic rules in data-driven models. The interesting objective to better understand the machine-based production process, or to achieve model improvements, is usually hidden, e. g., in the trained weight matrix of a neural network. To introspect data driven models, methods like classification and regression trees (CART) should be used alternatively.
The authors want to improve accuracy and perceptual acceptance of the IGM by applying evolutionary algorithms. By optimizing, e. g. minimizing, the neural network core and/ or the input vector, also the introspection of trained network configurations becomes easier. Finally, minimized structures support the integration of resource-saving prosody modules into embedded text-to-speech applications.

2. Integrated model of German prosody (IGM)

2.1. Training data

The analyzed data were part of a German corpus compiled by the Institute of Natural Language Processing, University of Stuttgart, and consists of 48 minutes of news messages from the radio station Deutschlandfunk, read by a male speaker. The database contains 356 sentences with 5.726 words including 13.151 syllables. The messages were recorded and partly repeated with an offset of 30 minutes. For the investigation 29.362 phonemes were available.
Regardless of a few word replacements, selected message texts are identically stored in the database. These messages basically differ in their recording dates. The data analysis considers also these recurrences.
This corpus does not contain spontaneous utterances. Both, news reading style and individual speaker characteristics are well-defined and reproducible. With regard to the prosodic target model for speech synthesis this reading corpus seems to be appropriate.

2.2. Model approach

Most conventional TTS systems calculate prosodic parameters sequentially, generating syllable durations first and then aligning the fundamental (f0) contour appropriately. This method does not sufficiently take into account that intonation and speech rhythm coherently exist.
The modeling of the production process of prosody and the interrelations between the prosodic features of speech are far from being a solved problem. Based on these considerations, the objective of Mixdorff and Jokisch was the development of a prosodic model taking into account the coherence between melodic and rhythmic properties of speech [3].
The model was henceforth to be called an 'integrated prosodic model', 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.

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.

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 [3].

Output parameter of model	Calculated as	N of tokens in database
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

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 training corpus. In order to capture potential interactions between intonation and rhythm, the prosodic parameters are predicted from a set of 24 linguistic and phonetic input features using a single, multi-layer feed-forward neural network (MFN, see Figure 1), 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.

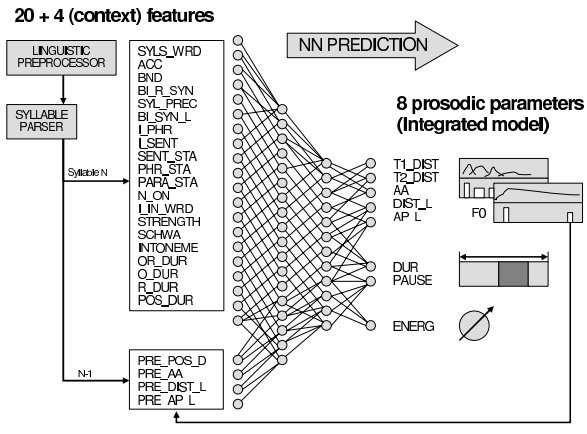

Figure 1: MFN structure (neural network core of IGM).

3. Limitations and problems using IGM

The IGM model uses an appropriate set of linguistic and phonetic input features (syntax, phrasing, accentuation, phone classes, etc.). The model was tested with resynthesis and synthesis stimuli [3] and its performance is suitable for several synthesis applications.
Nevertheless, the following factors influence and limit the overall performance:

3.1. Lack of semantic input information

The model does not consider elements of meaning, components like *semantic focus* or derived features like the roles of local prominence in disambiguation of focal adverbs, relative prominence in anaphora resolution and global prominence (register features) in the discourse structure.
Studying the correlation between prosody and semantics, the authors of [6] identified following discourse relevant register categories:

- Modifying features,
- Underlying features.

According to Discourse Representation Theory (DRT), they tested, whether the local prominence influences the meaning resolution by assuming three types of reading:

- First of sequence reading (FS),
- Exclusion of preceding alternatives reading (EPA),
- Retardation reading (R).

Training data and input vector of the IGM model need to be systematically extended by such semantic information.

3.2. Empirical neural network topology

The neural network topology of the IGM model was designed according to experiences from similar tasks in pattern recognition and prediction. There is no determinate rule system to find the optimal configuration (number of hidden layers and neurons, transfer functions, etc.).
The prediction performance of the final neural network is depending from well-defined training and testing sets, initial boundary conditions, learning rate and from other factors.
Under certain conditions, an evolutionary optimization can eliminate redundancies in the topology and it can also improve the significance, e.g., of remaining input features in further training cycles of the neural network.

3.3. Calculation complexity and memory consumption

Considering implementation issues, neural network structures like MFN usually provoke a higher demand for system resources than conventional rule-based models. Currently, IGM is not suitable for use in embedded text-to-speech systems.
Optimizing the empirical network topology and the dimension of the input vector can provide smaller model configurations, while widely keeping or even improving the prediction quality of the model.

4. Evolutionary optimization

Evolutionary algorithms (EA) may be described as stochastic optimization methods which simulate the process of natural evolution, feature selection and variation. Selection bases on the survival of the fittest. Individual solutions compete for

resources and reproduction. The recombination and mutation of genomes is called variation. EA are commonly used to solve problems in a wide range of fields including speech recognition and speaker verification.
In [7], Takagi gives a broad overview with regard to the use of interactive evolutionary computation (IEC) for optimizing systems based on subjective human evaluation as, e.g., in [8] and [9]. Beside IEC, there are surprisingly scarcely cases, in which EAs have been applied to prosodic issues. Recently, Kruschke proposed an evolutionary optimization for the automatic extraction of Fujisaki intonation parameters from a given speech database [10].

4.1. Multiobjective Optimization Problem (MOP)

Most of the real-world problems contain more than one objective. Optimizing more than one parameter causes an infinite number of optimal problem solutions, generally known as Pareto-optimal solutions. This set of solutions is also called the Pareto Front.
The concept of Pareto Dominance describes relationships between different solutions. One solution dominates another one, if exceeding it at least concerning a single parameter while no other parameter is worse. In Figure 2, solutions in the dark gray rectangle (bottom left) are dominated by *B*. But *B* itself is dominated by the rectangle upper right. Solutions *F* or *C* are neither dominated by *B* nor dominate *B*, although they are not optimal, as well. Solutions like *A*, along the Pareto Front, are optimal.

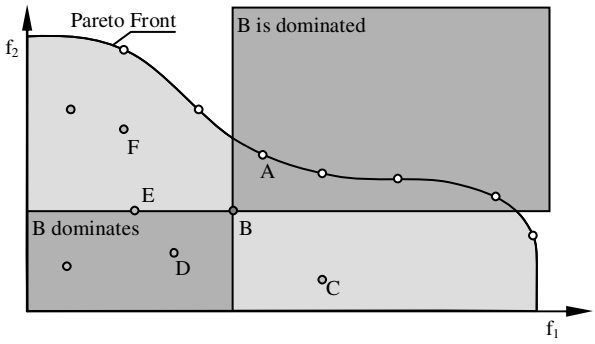

Figure 2: Pareto Front.

4.2. Strength Pareto Evolutionary Algorithm (SPEA)

To approximate the Pareto Front, different algorithms have been developed. Considering the multiple input parameters and the topology of the IGM model, the comparably new Strength Pareto Evolutionary Algorithm (SPEA) seems to be appropriate.
It externally stores all non-dominated solutions found so far as the *elite*. The fitness of an individual solution is only determined by the relationship to the others from the elite. The number of solutions of the elite is kept small by means of clustering, and dominated ones are removed.
The SPEA algorithm is described in [11], in more detail.

4.3. SPEA application to the neural network

The neural network design process is limited by two extremes: Either the network is so small that it is incapable of learning all training patterns and their dependencies, or it is too big to generalize the available data, and therefore it is learning each pattern separately.
There are three potential methods for the optimization of the described network core (MFN) in the IGM model:

1. The topology of the network,
2. Decreasing number of input parameters,
3. Increasing number of input parameters.

Third method would require a redesign of the whole model approach and is not pursued further in this paper. By using EA, there are the following optimization goals:

- Minimization of root mean square error (RMSE),
- Minimization of the number of links/ connections,
- Minimization of the number of inputs,
- Minimization of hidden layers.

To solve the mentioned multiobjective problem, the concept of Pareto Dominance is used.

4.4. Experiments and results

The original MFN consists of 24 inputs, 744 connections, 30 hidden neurons, and 8 outputs. The minimal, observed RMSE after training amounts to 0.139.
Method 1 - enlarging the MFN topology does not lead to significant results. Two bigger network configurations (40 and 50 hidden neurons) perform similar as the original net. Probably, the original topology is already task-appropriate.
Running SPEA to reduce the overall network topology by deleting connections and neurons (whereby all iterations are followed by new standard back-propagation training) shows no significant variation of the resulting RMSE. Considering the time consumption of a few hours per evolution generation, experiments were stopped after several hundred iterations.

Method 2 – decreasing the number of inputs is leading to significant results as shown in Figure 3. An example network with only six inputs achieves a RMSE of about 0.145 which is only 4.4 % worse than the original one.

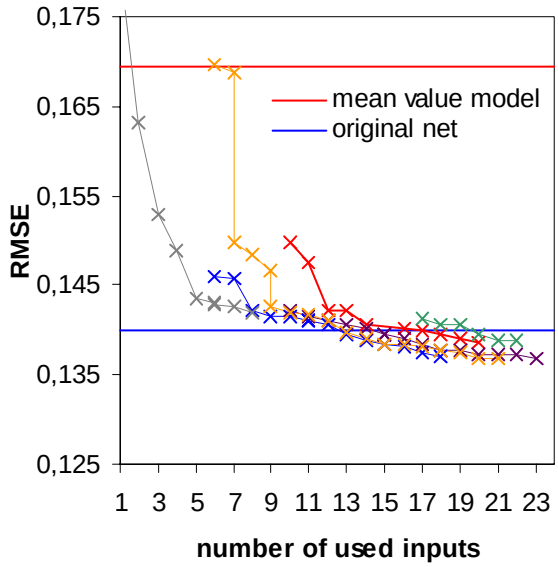

Figure 3: Several optimization runs compared to the original performance (bottom line) from [12].

The remaining six inputs after optimization are:

- *Bi_syn_r* (break index to the right),
- *Syl_prec* (syllable no. in preceding phrase),
- *Bi_syn_l* (break index to the left),
- *Schwa* (schwa vowel),
- *Intoneme* (accent type),
- *O_dur* (sum of mean phone duration in onset).

The complete input description of the MFN is given in [12]. Correlation analysis in [3] confirms that about 80-95 % of the prediction capabilities are basing on just 5-8 input parameters. Preliminary, subjective perception tests prove no significant audible differences among the resulting f0/ duration contours generated by different solutions.

4.5. Input data inconsistency

A further evolutionary experiment was addressing the correct selection of training and testing data. During the original training and adaptation of the IGM model, the recording conditions of the used part from the Stuttgart radio corpus were considered as almost constant during both sessions.
Since the EA showed significant preferences for special training and testing set combinations, an inconsistency of the input data was discovered. The database contains speech signals recorded on two different days. The waveforms of the first day seem to be normalized to the peak amplitude resulting in an average RMS power of about -17 dB. The waveforms of the second day were not modified and achieve an average RMS power of about -21 dB, a decrease of 4 dB.
This inconsistency explains different mean and standard deviation values of the parameter *Energy* (see table 2). The also varying values of the important output parameter Aa can be explained by different styles of speaking.

Table 2: Inconsistencies with regard to the input data.

Parameter	Mean	Standard deviation
Energy (day 1)	2437	1085
Energy (day 2)	1526	6198
Aa (day 1)	0.373	0.165
Aa (day 2)	0.2561	0.118

5. Conclusions

The proposed evolutionary optimization method SPEA seems to be an appropriate mean to verify empirical neural network configurations for the training of prosodic parameters. The vector dimension of a set of linguistic-phonetic input features can be reduced since SPEA is eliminating redundancies. So, this approach can reduce system resources, e.g. for designing embedded text-to-speech systems.
Furthermore, this evolutionary optimization method is able to highlight inconsistencies or contradictions in the training set and eases the potential introspection of implicit rules learned by the prosody model during adaptation.
Obviously, it was not possible to increase prediction accuracy or perceptive quality of the resulting synthetic speech.

Data-driven prosody models, such as IGM, need to be further qualified by additional knowledge sources, e.g., by semantic information or by language/ speaker-dependent data analyses (compare also IGM-related study on intensities [13]).
The proposed method can optimize the model structure but not the information content in the prosody model.

6. References

[1] W.N. Campbell, “Syllable-based segmental duration.” *In: Bailly, G. and Benoît, C. (ed.), Talking Machines: Theories, Models, and Designs*, 211-224, Elsevier Science, 1992.

[2] C. Traber, “F0 generation with a database of natural f0 patterns and with a neural network.” *In: Bailly, G. and Benoît, C. (ed.), Talking Machines: Theories, Models, and Designs*, 287-304, Elsevier Science, 1992.

[3] H. Mixdorff and O. Jokisch, “Evaluating the quality of an integrated model of German prosody”, *International Journal of Speech Technology (IJST) vol. 6 (issue 1)*, 45-55, Kluwer Academic Publishers, 2003.

[4] R. Hoffmann, D. Hirschfeld, O. Jokisch, U. Kordon, H. Mixdorff and D. Mehnert, “Evaluation of a multilingual TTS system with respect to the prosodic quality”, *Proc. ICPhS*, San Francisco, vol. 3, 2307-2310, 1999.

[5] O. Jokisch, H. Ding and H. Kruschke, “Towards a multilingual prosody model for text-to-speech”, *Proc. ICASSP*, Orlando, 421-424, 2002.

[6] G. Dogil, J. Kuhn, J. Mayer, G. Möhler, S. Rapp, “Prosody and discourse structure: Issues and experiments”, *Proc. ESCA Workshop on Intonation: Theory, Models and Applications*, Athens, 99-102, 1997.

[7] H. Takagi, “Interactive evolutionary computation: Fusion of the Capabilities of EC optimization and human evaluation”, *IEEE Proc., vol. 89, no. 9*, 1275-1296, 2001.

[8] Y. Sato, “Voice conversation using evolutionary computation of prosodic control”, *Proc. 12th Symposium on Human Interface*, 469–475, Yokohama, 1996.

[9] T. Morita, H. Iba, and M. Ishizuka, “Generating emotional voice and behavior expression by interactive evolutionary computation”, *62nd Annual Meeting of Japan Society for Information Processing*, 45–46, Yokohama, 2001 (in Japanese).

[10] H. Kruschke and A. Koch, “Parameter extraction of a quantitative intonation model with wavelet analysis and evolutionary optimization”, *Proc. ICASSP*, vol. 1, 524-527, Hong Kong, 2003.

[11] E. Zitzler, “Evolutionary algorithms for multiobjective optimization: methods and applications”, *PhD thesis, ETH Zurich*, 1999.

[12] F. Kossebau, “Evolutionary optimization of a trainable prosody computation, *Diploma thesis, TU Dresden*, 2003 (in German).

[13] O. Jokisch and M. Kühne, “An investigation of intensity patterns for German”, *Proc. EUROSPEECH*, 165-168, Geneva, 2003.