MULTI-LEVEL RHYTHM CONTROL FOR SPEECH SYNTHESIS USING HYBRID DATA DRIVEN AND RULE-BASED APPROACHES

Oliver Jokisch, Diane Hirschfeld, Matthias Eichner, Rüdiger Hoffmann

Technical Acoustics Laboratory, Dresden University of Technology,
D-01062 Dresden, Germany
Email: jokisch@eakss1.et.tu-dresden.de

ABSTRACT

This paper presents: a multi-level concept to generate the speech rhythm in the Dresden TTS system for German (Dre*SS*). The rhythm control includes the phrase, the syllabic and the phonemic level. The concept allows the alternative use of rule-based or statistical, but also data driven methods on these levels. To create the rules and to train a neural network, a new speech corpus from original speakers of the diphone-based inventories has been recorded. The corpus covers texts and single utterances and is subdivided into phrase, syllabic and phonemic databases.

First results indicating that the rule-based and the train-based methods generate a comparable speech rhythm, if the databases are uniform. The stepwise duration control on several prosodic levels shows promise as a method of producing a flexible rhythm depending on the specific TTS application.

1. INTRODUCTION

The control of the speech rhythm has an essential influence on the quality of synthetic speech. Beside common aspects of a duration control - like the correct modeling of segmental durations and a robust function - nowadays speech applications demand a higher rhythm flexibility (text reader, dialogue systems, etc.) and the realization of individual speaking styles.

With the higher segmental speech quality, also in the Dresden TTS system [1] for German, faults of non-acoustic processing stages, especially in the prosodic parts, are not longer masked. The redesign of the duration control postulates following theses:

- Global and local rhythm: The durations must be generated on several levels assuming the duration levels are not correlated.
- Availability of large databases and automatic analysis: Designing a prosodically-oriented database with respect to the synthesis target (re-implementing the individual rhythm of the inventory speaker, flexible speaking styles, etc.)
- Rule-based methods and data driven approaches can be combined.

According to the specific TTS application the duration control shall enable a "secure" speech output with a high intelligibility, respectively, e.g. an very "exciting" rhythm, which may contain mistakes, too.

The chapter 2 briefly discusses a multi-level approach for integrating global and local rhythm effects, while chapter 3 is describing the algorithms on these levels and the combinations. The necessary database - the new speech corpus - is explained in chapter 4.

2. THE MULTI-LEVEL RHYTHM MODEL

Most common for rule-based models is the direct estimation of the segmental target duration by applying linguistic and prosodic input features, which modify any base duration. Such as the Klatt model [2] perform a considerable standard rhythm on the phonemic level. Limitations occur, if the global speech rate will be strongly modified or sharp local accents are intended.

The multi-level approach supports the global and local flexibility and follows a top-down strategy including the phrase, syllabic and phonemic level (figure 1).

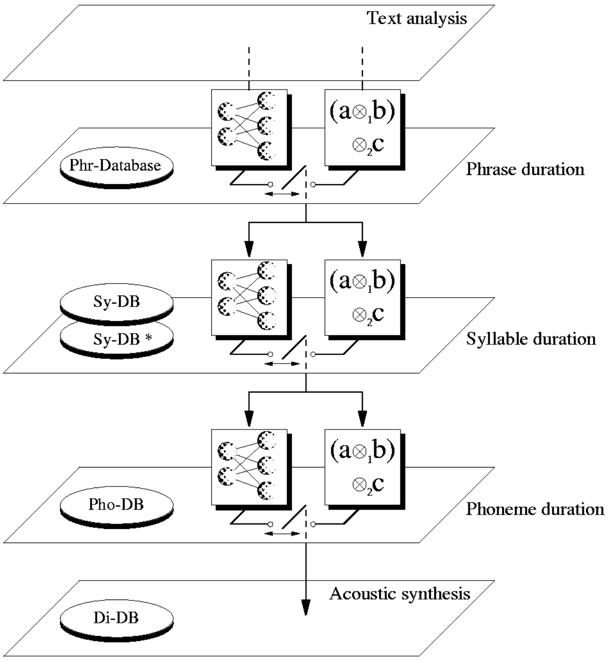

Figure 1: Multi-level rhythm approach: alternative neural network or rule-based methods for estimating the durations and separate databases on each level. Databases (DB) see chapter 4: Phr .. phrase, Sy .. syllable, Pho .. phoneme, Di .. diphone Sy-DB .. NCO syllables, Sy-DB* .. ONC syllables.

The hybrid design enables the flexible use of either a neural network (ANN) or a rule-based duration control algorithm on each level. During the training of the ANN and also for adjusting the rules - the input symbols like stress and phrase markers or phonetic attributes, but also the target duration will be separately applied to each level. This redundancy among the presented attributes theoretically allows the coexistence of different approaches for the duration control on the mentioned levels (e.g. the Klatt model in contrast to the syllable-oriented model from Campbell [3]).

To examine several paths through the multi-level module, following steps seem to be appropriate:

1. Audio-visual tests and manual adjustment of the single levels
2. The comparison between generated and target durations (with reference to the data analysis of the new speech corpus).
3. Perceptive preference tests

For the first step a graphical user interface was developed, which integrates neural and rule-based algorithms for the intonation and duration control. This interface provides comfortable facilities for adjusting parameters (speech rate, f0 baseline, dynamic, limiters, etc.) and for monitoring results (Neu*Rosy* [4]). The evaluation of the objective duration criteria and the perceptive results are described in [5].

3. METHODS

The TTS system currently contains three (alternative) approaches for the duration control: a rule-based, phoneme-oriented model (German adaption according to Klatt), another rule-based, phrase/syllable-oriented model (section 3.1), but also an ANN approach (section 3.2), which is mainly syllable-oriented. The ANN module has been designed with respect to an analogous intonation network in [4], which was influenced by Traber's intonation concept [6].

Both, the new rule-based and the ANN modules are using the statistical parameters of the new corpus described in chapter 4.

3.1. Rule-Based Duration Control

The rule-based duration control uses a set of rules for each level. These rules are extracted by a statistical analysis of the corpus. The phrase layer determines the duration for a given prosodic phrase depending on the number of syllables and the type of the prosodic phrase (equation 1). The phrase type is explained in section 4.2.

$$(1) \quad dur_{phr}(n) = k_i n + c_i \qquad \begin{array}{|ll} dur_{phr} & \text{duration} \\ n & \text{number of syllables} \\ k_i & \text{factor} \\ c_i & \text{constant} \\ i & \text{phrase type} \end{array}$$

The second layer calculates the duration of each syllable as a linear combination of the number of phonemes. Various phonetic attributes like accent or nucleus type influence the duration of the syllable in different ways; e.g. the nucleus type causes a lengthening by a factor and the accented syllable is stretched by adding a constant (equation2).

$$(2) \quad dur_{syl} = \begin{cases} \overline{dur_{syl}}\, k_{initial} & : \textit{initial syllable} \\ \overline{dur_{syl}} + c_{acc} & : \textit{accented syllable} \\ \ldots & : \ldots \end{cases}$$

Afterwards, the syllables are adjusted by linear stretching or shrinking in order to fit into the duration frame calculated on the phrase level.

Finally, the duration of each phone is adapted to the frame of the syllable duration according to Campbell's elasticity hypothesis: A stretching factor (z-score) is iteratively calculated for a given syllable duration and the means and standard deviations of the phoneme durations. The z-score is assumed to be constant all over the syllable. The actual phone duration is calculated from the z-score and the mean and standard deviation of its specific phoneme duration.

3.2. ANN Approach

The general concept for the training of the target durations is presented in figure 2 and can be implemented on each level.

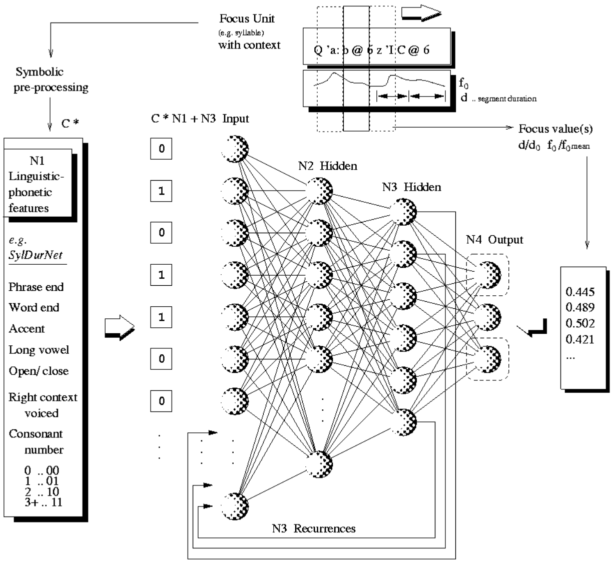

Figure 2: ANN model for the duration (and intonation) control. The sample of input features is taken from the syllable level.

The following paragraph describes the process on the syllable level: From the phonemic text syllables will be isolated and stepwise presented to a recurrent network (including 1 "focus syllable" and 2 syllables "context": context frame C=5). For each syllable a vector of N1=8 linguistic and phonetic features is applied to the network input. The first hidden layer consists of N2=10, the second hidden layer of N3=6 neurons. The second hidden layer is completely connected to the "context neurons", i.e. the ANN input layer contains C*N1+N3=46 neurons. The output layer owns only N4=1 neuron, which estimates the syllable duration of the "focus syllable". This hierarchical Elman network is trained with an adapted error-backpropagation algorithm using the distances between original durations from the speech corpus and the net output.

The input coding considers the phrase position and stress situation, the phonetic attributes of the nucleus and its context. Furthermore, the number of consonants in the syllable is coded (See also figure 2). Beside the appropriate coding, the network performance mainly depends on the train utterances or the selected text.

3.3. The Hybrid Module

With regard to the duration models presented, several combinations of methods are possible. Table 1 shows the (hybrid) architectures which have been already tested

Architecture	Phrase level	Syllabic level	Phonemic level
Single-level rule (SLR)	-	-	Klatt
Multi-level rule (MLR1)	RB	RB (ONC)	Campbell
Multi-level rule (MLR2)	RB	RB (NCO)	Campbell
Multi-level net (MLN)	ANN	ANN (NCO)	ANN
Hybrid rule/net (HRN)	RB	ANN (NCO)	Campbell
Hybrid net/rule (HNR)	ANN	RB (ONC)	Campbell

Table 1: Overview about the hybrid configurations that are available across the multi-level module:
RB .. rule-based (see 3.1),
ANN .. artificial neural network (see 3.2)
ONC .. onset-nucleus-coda (phonologic syllable, see 4.1)
NCO .. nucleus-coda-onset ("pseudo" syllable, see 4.1)

Phonemic level. For the duration distribution on the phonemic level, originally, the phoneme-oriented approach (Klatt) and the ANN method were implemented. Because of the proper results given by the elasticity-"z-score"-model, for testing the MLR1, MLR2, HRN and HNR modules, Campbell's approach was used.

4. SPEECH CORPUS

4.1. Database Design

In order to adjust the rule-based and the statistical algorithms but also to train the neural networks - new speech data from both, male and female, original speakers of the diphone inventory have been recorded. The data of our male speaker (native speaker of German, f0=100Hz) can be subdivided into two parts:

1. The text corpus (344 sentences, 10780 segments) was selected to show natural prosodic effects and a speech rhythm typical for a text reading application. It combines two short stories and a longer passage of a coherent text from a story tale.

2. For the purpose of inventory extraction the sentence corpus (443 sentences, 11353 segments) contains all phoneme combinations in the German language. On the other hand, the demand for a natural and fluent speaking style requires the embedding of the units into a sentence context. Both demands are met by the recorded sentences similar to the German PhonDat 1 - corpus [7].

Data preparation. The natural speech signal was labelled using information from different linguistic description levels. Much attention was paid to provide labels on the base of objective features. The labels should be re-useable for other purposes (training of automatic labellers, inventory generation, statistic studies, etc.).

Phone labels. The SAMPA-inventory for German was extended by symbols, e.g. for pauses, noise and segments to be excluded from further processing. Plosives were subdivided into two segments: pause and burst including aspiration phase. The labelling of vowels was done on the base of formant features [8].

Prosodic labels. The labelling of accents (PA, WA) and phrase boundaries was done on the base of smoothed z-score-traces and pitch contours (See figure 3).

Syntactic labels. Finally, labels for syllabic, word and clause boundaries were manually provided.

Syllable types. Two alternative definitions of the syllable were used for pragmatic reasons: The NCO-"pseudo" syllable is enclosed by two vowels. The syllable starts at a vowel and ends before the next vowel. That keeps the syllabification process simple. Word boundaries are not included in the hierarchy built up by NCO-syllables. The next higher level is the phrase or clause. At the phrase begin, there are rudimentary syllables without vowel, that are excluded from further processing.

The second type, the ONC-syllable, is oriented on phonologic/ acoustic criteria. Word boundaries and prefix or suffix boundaries are matching the syllabic boundaries. The position of syllabic boundaries in consonant clusters considers the acoustic segmentation of the speech signal (within plosive stops/ after voiceless fricatives). To prevent open syllables containing short vowels, single inter-vocalic consonants are distributed to both neighbor syllables.

4.2. Data Analysis

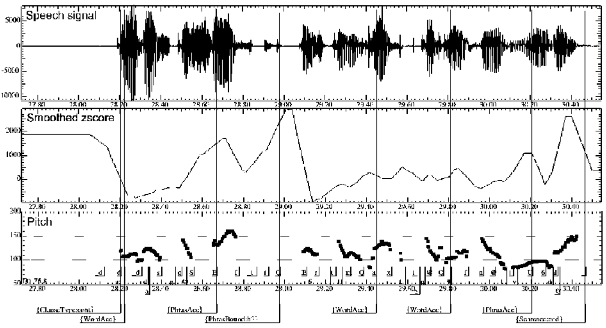

Figure 3: Example - data analysis by using smoothed phoneme z-scores and prosodic labeling

For the analysis of phrase, syllabic and phonemic durations all prosodic and syntactic label files were projected to the phone labels. All relevant information was extracted automatically: For the raw duration distribution of each phoneme, mean and standard deviation were calculated. To compensate the skew of the distributions the logarithm of the raw duration was taken into account.

Beside the phonemic database (Pho-DB), for both syllable types (NCO, ONC) a syllabic database (Sy-DB) was constructed containing the following information: Index of syllable in the word, index of word in the clause, index of clause in the speech file, filename, phoneme string, duration of the syllable, nucleus type (long vowel, short vowel, diphthong, reduced vowel and syllabic consonant), accent type, function word, phrase- and word position (initial, medial, final), number of phonemes and relative position of the nucleus.

The phrase database (Phr-DB) was constructed to contain *prosodic* phrases. It contains information about: index of the clause in the speech file, filename, phrase duration, phrase type and the number of syllables in the phrase. The following *phrase-types* are examined: "clause begin - word accent", "clause begin - phrase accent", "word accent - phrase accent", "word accent - word accent", "phrase accent - clause end".

5. CONCLUSION

The presented multi-level approach simultaneously supports the processing of global and local factors on the segment duration, keeping the duration module in an uniform design and providing:

- Rule-based, statistical or ANN methods and their combinations
- A generalized database for all methods and the acoustic synthesis

The introduced methods differ concerning the effort of rule adjustment, train data collection, etc.. Nevertheless, the generated target durations are quite similiar on each level.

The statistical approach (Campbell), produces proper results on the phonemic level. On the syllabic level the phonolgic ONC-syllable is appropriate for both, rule-based and ANN, methods. The numeric and the perceptive results are presented in [5].

The main advantage of the approach results from the speaker-specific data with regard to the prosodic database and the acoustic inventory. The further work will focus on the phrase level optimization. It is intended to enlarge the phrase corpus and the set of prosodic features.

6. REFERENCES

1. Hirschfeld, D., Maas, H. D.: Improving the functionality of a text-to-speech system by adding morphological knowledge. Proc. 20th Annual German Conference on Artificial Intelligence (KI96), Dresden, 103-106, 1996.
2. Klatt, D. H.: Review of text-to-speech conversion for English. J. Acoustic. Soc. Am., 88: 737-793, 1987.
3. Campbell, W. N., Isard, S. D.: Segment durations in a syllable frame. J. of Phonetics , 19: 37-47, 1991.
4. Jokisch, O., Pescheck, M.: Neuronale Prosodiegenerierung - Einfluss der Trainingsdaten (in German). Proc. 24th Annual German Conference on Acoustics (DAGA), Zurich, 1998 (in print).
5. Jokisch, O., Hirschfeld, D., Hoffmann, R.: Creating an individual speech rhythm: a data driven approach. Proc. ESCA Workshop on Speech Synthesis, Jenolan, Australia, 1998 (in print).
6. Traber, C.: F0 generation with a database of natural f0 patterns and with a neural network. In G. Bailly, C. Benoit, ed., Talking Machines: Theories, Models, and Designs, 287-304, North Holland, Amsterdam, 1992.
7. PhonDat 1, BAS corpora on CD-ROM, Institute of Phonetics and Speech Communication, Munich.
8. Hirschfeld, D.: Variabilitaet und Stabilitaet segmentaler Merkmale unter dem Aspekt der konkatenativen Sprachsynthese – Vokale (in German). Proc. 7th Conference on Electronic Speech Signal Processing (ESSV), Berlin, 94 - 101, 1996.