EVALUATION OF A MULTILINGUAL TTS SYSTEM WITH RESPECT TO THE PROSODIC QUALITY

Rüdiger Hoffmann, Diane Hirschfeld, Oliver Jokisch, Ulrich Kordon,
Hansjörg Mixdorff, and Dieter Mehnert
Technische Universität Dresden, Institut für Akustik und Sprachkommunikation

ABSTRACT

Improving the naturalness of synthetic speech is an essential task in developing a text-to-speech (TTS) system. Mainly, it depends on the quality of the prosody model which is utilized in the TTS system. For our TTS system called DreSS (Dresden Speech Synthesizer), we compared three different methods for generating the F0 contour to each other as well as to other synthesizers. Natural speech samples were used as a reference. Results show, that on a naturalness scale from 0 to 4, the natural speech samples reach a maximum score of 3.6, with values of 1.9 for the best synthesis, the LPC-based one. The system with an intonation control basing on the Fujisaki model leads the group of PSOLA systems, which are closely clustered at a mean of 1.54.

1. INTRODUCTION

This paper describes recent improvements of the Dresden TTS system (DreSS) and their evaluation. DreSS is a diphone-based time-domain synthesizer with preprocessing module, grapheme-phoneme converter, duration control, intonation control, and acoustic module. It is available as a software system but also as stand-alone system supported by a special processor [1]. This solution is available with PCMCIA standard.

Recent improvements refer to multilinguality and naturalness. Multilinguality is obtained by a dedicated structure which can handle databases from different languages. Databases for German, English, Russian, Czech, and Chinese have been developed.

Naturalness is a highly important feature of synthetic speech. Apart from the segmental quality and the voice characteristics, it depends mostly from the prosody. Because it is hard to evaluate in an objective way, we started a perceptual comparison of different methods for generating the F0 contour of German sentences. For this purpose, DreSS was equipped with three intonation modules (rule-based linear f0-model, neural-network based approach, rule-based approach applying the Fujisaki model). In this paper, we describe the intonation modules as well as the results of the evaluation.

2. THE TTS SYSTEM DreSS

This chapter gives an overview of the text processing in the **Dre**sden **S**peech **S**ynthesis System (DreSS) as shown in Figure 1:

- Plain ASCII-text or text enriched with conceptual information - containing pronunciation forms of some words, accent- or boundary-tags - is processed by the preprocessor stage. Word, phrase, and sentence boundaries are classified and tagged, special character combinations are recognized, function words and abbreviations are detected and marked in the running text.

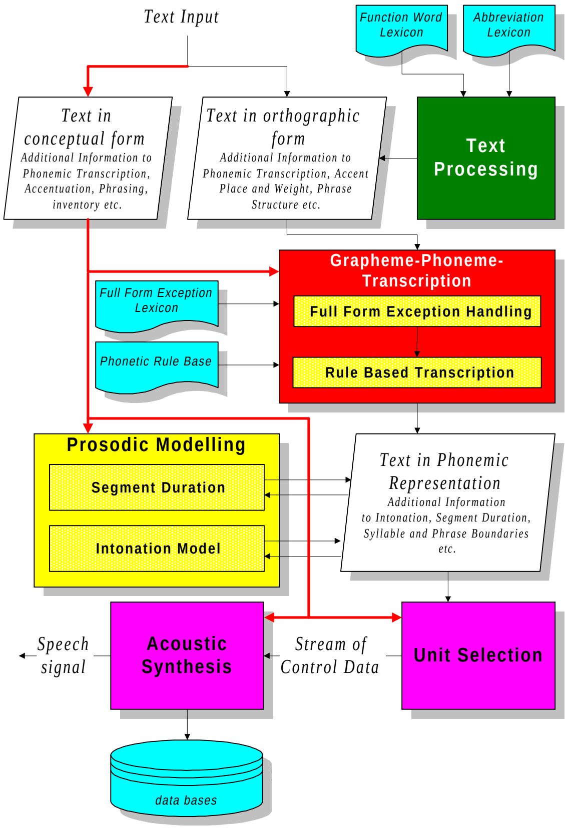
Text Input
Function Word Lexicon
Abbreviation Lexicon
Text in conceptual form
Additional Information to Phonemic Transcription, Accentuation, Phrasing, inventory etc.
Text in orthographic form
Additional Information to Phonemic Transcription, Accent Place and Weight, Phrase Structure etc.
Text Processing
Grapheme-Phoneme-Transcription
Full Form Exception Lexicon
Full Form Exception Handling
Phonetic Rule Base
Rule Based Transcription
Prosodic Modelling
Segment Duration
Intonation Model
Text in Phonemic Representation
Additional Information to Intonation, Segment Duration, Syllable and Phrase Boundaries etc.
Speech signal
Acoustic Synthesis
Stream of Control Data
Unit Selection
data bases

Figure 1. Processing stages in DreSS.

- The Grapheme-to-Phoneme stage derives a phonemic representation of the input text. Processing is done first by a lexicon-based and then by a rule based component. Furthermore, accent type and place are supplied to the following prosodic components.
- Prosodic processing (duration and intonation control) is done by several modules, between which the user may choose. The different approaches, which will be discussed in more detail in the

following chapters, add segmental durations and pitch parameters to the stream of phonemic information.

- The unit selection transforms the stream of phonemes into a sequence of speech units (Diphones) and joins it to the prosodic information.
- Finally, the acoustic synthesis builds up a synthetic speech signal from the sequence of Diphones and reproduces the prosodic parameter contours.

3. THE INTONATION MODELS

3.1. Linear Approach

The linear approach to intonation control is a very simple production model. It superposes the contributions of accentuation-based and sentence/phrase-based linear components to a resulting intonation contour. In Figure 2, a short example is given.

Parameters like declination, accentuation rise and the phrase components may be configured by the user.

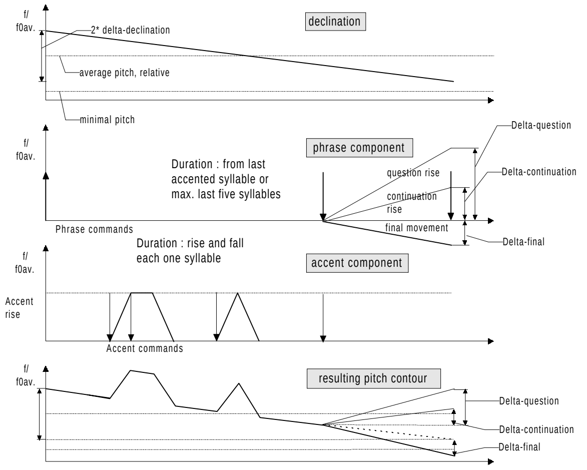

Figure 2. Intonation control with the linear approach.

3.2. Data Driven (Neural) Approach

To generate flexible speaking styles and to quickly adapt the DreSS system to the requirements of different voices or languages – a data driven approach is used. This approach includes a artificial neural network (ANN) and enables the direct estimation of the f0 contour from a sequence of linguistic input vectors. The feature coding is syllable-oriented: From the phoneme sequence, syllables will be isolated and stepwise presented to a recurrent network (including a focus syllable and in each case two syllables for the pre- but also for the post-context, which means a context frame of 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 N4 = 3 neurons, which estimate the f0 contour of the focus syllable. The input encoder considers the phrase position, stress situation, phonetic features of the nucleus and its context (see Figure 3).

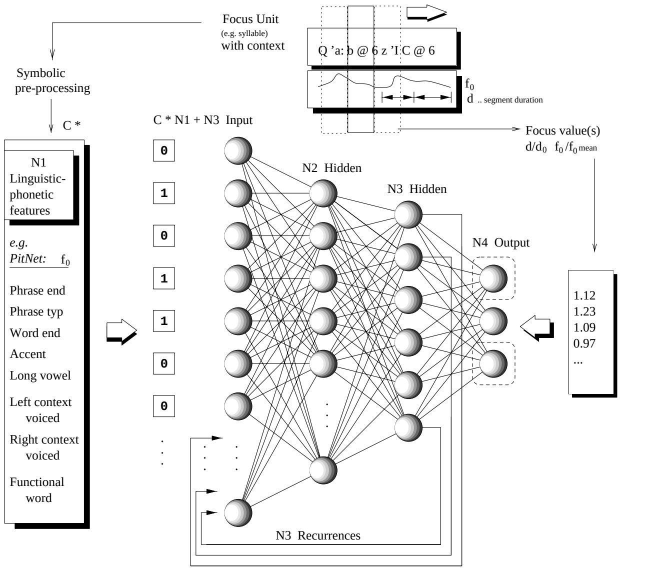

Figure 3. ANN model for the intonation control. (An analogous approach for the duration modeling is additionally shown.)

The hierarchical Elman network is trained with an adapted error-backpropagation algorithm using the distances between the original f0 contour extracted from the training corpus (PhonDat I [2] - up to 400 sentences of a single speaker) and the net-generated f0 contour. Beside the appropriate coding, the network performance mainly depends on the amount of training utterances and the choice of the presented examples. The training corpus for the ANN model used for the presented prosodic evaluation was not recorded from source speakers of the DreSS voice inventory. Another study [3] deals with the potentials of matching databases.

3.3. Adaptation of the Fujisaki Model

The third method utilized in DreSS for generating the intonation applies the well-known Fujisaki model (e.g., [4]). This model represents the combination of a phrase control mechanism (controlled by phrase commands) and an accent control mechanism (controlled by accent commands) according to a mathematical algorithm. Since this algorithm considers facts from the articulatory apparatus, the model is suited for different languages. Of course, the parameters of the model have to be determined for every language. For German, this was done by one of the authors in [5] and [6]. Thus, we are able to generate the accent and phrase information for the Fujisaki model from a given sentence in DreSS. In this paper, the method is referred to as MFGI (Mixdorff Fujisaki German Intonation).

4. DURATION CONTROL

The presented evaluation of the prosodic quality mainly examines the effect of different intonation models on the overall rating of a TTS system. The control of segmental durations has an essential influence on the quality of the synthetic speech, too. For the internal evaluation of several DreSS intonation models - but also for the standard use of the TTS system - a rule-based duration control according to the approach of Klatt [7] is implemented, because of its robust performance and satisfying mean quality: Basing on the phonemic word string and the obtained accent level (0, 1, 2), the duration control generates a duration for each phoneme using a rule system while modifying inherent and minimal durations from a given phoneme database.

Other than the phoneme-oriented rule model, a new multi-level approach [8] supports a global and a local rhythm variation and follows a top-down strategy including the phrase, syllabic and phonemic level. The concept allows the alternative use of rule-based, statistical or data-driven methods on these levels. The hybrid design enables the flexible use of either a rule-based duration control or a neural network control (see the analogous model in Figure 3) on each level. However, this model was not evaluated in the study presented here.

Summarizing, the generated durations do not differ significantly among the rule model and the ANN approach on the phrase and the syllabic levels - as long as the database for the training and/or the rule adjustment are identical. For the phonemic level, a statistical model with regard to Campbell's elasticity approach (e.g. in [9]) showed the best results.

5. EVALUATION

5.1. Test Material and Persons

To compare the different methods for generating the intonation on *sentence level*, we produced some examples for terminal, progredient, and interrogative sentence intonation. For the perception experiment on *system level*, a continuous text was selected from news material.

53 listeners had been available. Among them, 21 were trained (phonetically educated and familiar with synthetic speech), while the other 32 listeners had no contact to synthetic speech formerly.

Table 1 gives an overview on the experiments. It is shown that we performed at first two blocks of experiments where we compared the different intonation modules in DreSS [10]. A version of DreSS which showed good quality in these experiments, was finally compared to other TTS systems and to natural speech in a third block [11].

5.2. First Experiment: Comparison within DreSS

Because a test of prosodic quality will be only reasonable if the synthetic speech shows a good basic quality, we tested at first the intelligibility of the synthetic sentences. All four methods mentioned in Table 1 produced a sentence intelligibility between 92 and 100 %.

Next, the same sentences were used to determine the naturalness. The listeners had to perform an absolute category rating (ACR) in a four step scale. MFGI performed best with a score of 1.7. The worst score (for NN II) was 2.7. Thus, the difference between the four methods was only one point.

A third part of this block concerned the perception of accents in the synthetic sentences. The listeners used a printout of the sentences to mark by a slash the places where they perceived an accent. The percentage of these marked accents compared to the intended accents was calculated. MFGI showed clearly the best result with a mean of 86.9 %. An overview (mean of all listeners) is shown in Table 2.

To complete the first block, the naturalness of the three best approaches was evaluated in a software-supported pair comparison. A score for the naturalness was calculated as mean of all points which were collected in all pair comparisons over all listeners. From this calculation, a scale from 0 to 4 results. Table 3 shows the results which are confirming the ranking resulting from the previous listening tests.

Table 1. Overview on the perception experiments.

	1st experiment				**2nd experiment**	**3rd (main) experiment**	
Comparison	DreSS internal					different TTS systems	
Feature	Intelligibility	Naturalness (ACR)	Marking of accents	Naturalness (A/B)	Naturalness (A/B)	Naturalness (A/B)	Naturalness (Ranking)
Compared methods and systems	DreSS with following intonation modules: • MFGI (see 3.3. above) • Neural network (NN) I (see 3.2. above) • Neural network II (smaller training set than NN I) • Linear approach (see 3.1. above)				• MFGI • NN (modified) • Linear • Copy contours • Copy contours with natural durations	• MFGI • NN (modif.) • Alien A • Alien B • Alien C • Natural speech	• MFGI • NN (modif.) • Alien A • Alien B • Alien C • Natural speech • MFGI with natural dur.
Test material	72 sentences	72 sentences	72 sentences	14 sentences (84 pairs)	8 sentences (125 pairs)	15 sentences (215 pairs)	3 connected sentences
Method	Opinion test	Opinion test	Opinion test	Pair compar.	Pair comparison	Pair compar.	Ranking
Result	Intelligibility in all methods nearby 100 %	1. MFGI 2. Linear approach 3. NN I 4. NN II	1. MFGI 2. Linear approach 3. NN I 4. NN II	1. MFGI 2. Linear approach 3. NN I 4. NN II	1. Copy contours/ natural dur. 2. MFGI 3. Copy contours 4. NN (modified) 5. Linear approach	1. Natural speech 2. Alien B 3. MFGI 4. Alien C 5. Alien A 6. NN (mod.)	1. Natural sp. 2. MFGI with natural dur. 3. MFGI 4. Alien C 5. NN (mod.) 6. Alien B 7. Alien A

Table 2: Perception of intended accents.

Method	Mean [%]	Standard deviation
MFGI	86.9	4.3
Linear approach	67.9	9.5
NN I	57.1	7.3
NN II	49.6	2.9

Table 3: Score of naturalness in pair comparison (Scale 0 ... 4).

Method	Mean	Standard deviation
MFGI	2.34	0.31
Linear approach	1.59	0.41
NN I	1.45	0.53

5.3. Second Experiment: Comparison to Natural Contours

In another pair comparison, we tried to compare the quality of the synthetic F0 contours to natural contours. For this purpose, the contours from natural utterances had been copied to synthetic stimuli. Additionally, a second set of these stimuli was equipped with natural durations.

The ranking resulting from these experiments is shown in the corresponding column of Table 1. Obviously, approaching natural sound durations leads to an essential improvement. This means that a good F0 control will be effective only if the duration control shows comparable quality.

5.4. Third Experiment: Comparison to Other TTS Systems and to Natural Speech

The purpose of this (main) experiment was to compare DreSS (equipped with MFGI as the F0 control which worked best in experiments 1 and 2) to other TTS systems and to natural speech. For this purpose, utterances from three renowned TTS systems for German had been produced via web access. The systems will be called here ‘Alien A, B, C’. Two of them are PSOLA systems, the third utilizes LPC segments.

The experiment was subdivided in two parts.

The first part was a pair comparison of isolated sentences analogous to the previous experiments. The results are shown in Table 4. Of course, there is still a remarkable difference between natural speech and the best TTS system.

Table 4. Overall rating of the test sentences (scale 0 ... 4).

System	Mean	Standard dev.
Natural speech	3.35	0.18
Alien B	1.88	0.42
DreSS with MFGI	1.65	0.29
Alien C	1.49	0.49
Alien A	1.48	0.32
DreSS with Neural Network	1.23	0.40

The second part was a system ranking. The aim was to measure the acceptance of a TTS system at the listener. The pair comparisons shows that the rating of the naturalness strongly depends on the sentence used. Furthermore, TTS systems are generally used for synthesizing connected texts. That’s why, a connected text from three sentences (from news) was selected for synthesizing by the systems to be compared. Additionally, we tested a MFGI version with natural durations. The resulting ranking was slightly different to that from the pair comparison and is shown in Table 5.

Table 5. Comparison of TTS systems (scale 0 ... 10).

System	Mean	Standard dev.
Natural speech	10.00	0
DreSS with MFGI and natural durations	6.57	2.02
DreSS with MFGI	5.35	2.60
Alien C	4.83	2.55
DreSS with Neural Network	4.70	2.58
Alien B	4.43	2.69
Alien A	2.65	2.74

6. CONCLUSION

As a result of our evaluation, we found a version of DreSS which compares well to other leading TTS systems for German. On the other hand, it proved again that recent TTS systems in general are far from the quality of natural speech. The improvement of the duration control seems to be an essential part to further enhance the quality of DreSS. Readers who are interested in testing DreSS are referred to our tutorial web pages [12] with the web address http://www.ias.et.tu-dresden.de/kom/lehre.

ACKNOWLEDGMENT

This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG), grant no. Ho 1673/2-1.

REFERENCES

[1] Hoffmann, R., Kordon, U., Holland, H. J., Netz, S. 1999. Multilingual speech synthesis for car applications. *Proc. ISATA ’99*, Vienna, paper 99AE003.

[2] PhonDat 1, *BAS corpora on CD-ROM*, Institute of Phonetics and Speech Communication, Munich.

[3] Jokisch, O., Hirschfeld, D., Eichner, M., Hoffmann, R. 1998. Creating an individual speech rhythm: a data driven approach. *Proc. 3rd ESCA/ COCOSDA Workshop on Speech Synthesis*, Jenolan, Australia, 115-119.

[4] Fujisaki, H. 1997. Modeling the process of fundamental frequency control of speech for synthesis of tonal features of various languages. *1997 China-Japan Symposium on Advanced Information Technology*, Invited Plenary Lecture, Anhui, 1 - 12.

[5] Mixdorff, H., Fujisaki, H. 1995. A scheme for a model-based synthesis by rule of F0 contours of German utterances. In *Proceedings of the '95 Eurospeech*, Madrid, Spain, vol. 3, 1823-1826.

[6] Mixdorff, H. 1998. *Intonation Patterns of German – Quantitative Analysis and Synthesis of F0 Contours*. Ph.D. thesis, TU Dresden, Dresden, Germany.

[7] Klatt, D. H. 1987. Review of text-to-speech conversion for English. *J. Acoustic. Soc. Am.*, 88: 737-793.

[8] Jokisch, O., Hirschfeld, D., Eichner, M., Hoffmann, R. 1998. Multi-level rhythm control for speech synthesis using hybrid data driven and rule-based approaches. *Proc. ICSLP ’98*, Sydney, 607 – 610.

[9] Campbell, W. N., Isard, S. D. 1991. Segment durations in a syllable frame. *J. of Phonetics*, 19: 37-47.

[10] Mixdorff, H., Mehnert, D. 1998. Perceptual evaluation of three different approaches for generating F0 contours in TTS. In *Fortschritte der Akustik 1998*, Zürich, Switzerland, 398-399.

[11] Mixdorff, H., Mehnert, D., Hirschfeld, D. 1999. Comparing the naturalness of several approaches for generating f0 contours in German text-to-speech systems. *Proc ASA ’99*, March 1999, Berlin, paper 4pSCa.

[12] Hoffmann, R., Kordon, U., Kürbis, S., Ketzmerick, B., Fellbaum, K. 1999. An interactive course on speech synthesis. *Proc. MATISSE ’99*, London, paper 016.