# Perceptual Comparison of Three Different Approaches for Generating F0-Contours in Text-to-Speech

Hansjörg Mixdorff and Dieter Mehnert

Institut für Technische Akustik, TU Dresden, Mommsenstr. 13, 01062 Dresden

## 1. INTRODUCTION

In German, as in many other languages, the $F_0$ contour is an important acoustic correlate of prosody. Hence, generating near-to-natural $F_0$ contours is necessary for achieving better intelligibility and naturalness of synthetic speech.

Comparing the prosodic quality of individual TTS-systems is very difficult, as the quality of linguistic analysis, the segmental quality and voice characteristics vary considerably.

For this reason, the current study examines the quality of three different approaches for modeling $F_0$-contours for speech synthesis, which have been integrated into the same TTS-system. We compare the rule-based linear model by Hirschfeld [1], the rule-based approach by Mixdorff/Fujisaki [2], and the neural-network (NN)-based approach by Jokisch [3], all of which have been developed at TU Dresden.

The corpus used for the study contained 39 sentences of statements and questions of varying length.

## 2. THE TTS-SYSTEM USED FOR THIS STUDY

The stimuli for the experiment were produced with a modified version of the diphone-based PSOLA-speech synthesizer at TU Dresden [1]. Figure 1 displays a block diagram of the system components. First, the text passes a **pre-processing module**, where it is split into phrases, which are typically delimited by colons or other punctuation marks.

The **grapheme-phoneme conversion module** (GPC) converts each word of a phrase into a corresponding SAMPA-string plus part-of-speech (p-o-s) and word accent information. In the **phonetic module**, which we will discuss a little more in detail, inter-word and phrase contexts are evaluated in order to derive accent levels for every word syllable. Long phrases are further split.

The **duration control module** generates durations for every phone in a phrase.

The **intonation module** produces a phone-aligned $F_0$ contour. As mentioned above, three versions of this module were tested: the rule-based approaches by Hirschfeld and Mixdorff/Fujisaki, and the NN-based approach. The model by Hirschfeld is the version originally used in the TU-Dresden TTS-system.

The last module in the synthesis chain is the **TD-PSOLA synthesis module**.

The linguistic rules applied in the phonetic module are based on Stock's and Zacharias' work on German intonation [4].

In the module, long phrases with more than 12 syllables are split into smaller parts. As there may be content words which are not contained in the dictionary and whose p-o-s is hence unknown, a complete syntactic analysis of a phrase may be impossible. For this reason a detection of function words, which are fully covered in the dictionary, is used in order to find additional prosodic phrase boundaries.

Consider the German sentence: "Er verfolgte seine ehrgeizigen Ziele ohne die geringste Rücksicht auf seine Familie."—*He pursued his ambitious aims without the slightest consideration for his family.* which contains 26 syllables. The phonetic module will detect the preposition-article sequence 'ohne die' and insert an additional phrase boundary before it.

Depending on its p-o-s, every constituent word is assigned an accent level. In principle, function words such as articles, pronouns, auxiliary words remain unaccented. All other words (nouns, verbs etc.) and also unknown words, receive a melodical accent on their word accent syllable.

In certain contexts, melodical accents are deleted, for instance on verbs in some noun-verb-sequences: "mit dem ˈAuto fahren."-*"to go by car."*

In phrases consisting of unaccented words only, an accent is added on the last auxiliary verb present "wir haben das gehˈabt." *"We have had that."*

## 3. PERCEPTUAL EVALUATION

The following perceptual experiments were conducted: (1) Assessment of naturalness (A/B-comparison), (2) Assessment of naturalness (scoring), (3) Marking of accent locations.

For the sake of conciseness we concentrate on some results from experiment no.1, a computer-based A/B-comparison of stimuli produced for a subset of 14 sentences. 22 subjects (12 male, 10 female) were asked to select the stimulus they found most natural in every pair of stimuli. All pairs were automatically randomized and presented twice. The subjects could listen to every stimulus as often as they liked.

In Figure 2 naturalness scores for all three kinds of stimuli are displayed with respect to the 14 test sentences. The scores are averaged pseudo-measures derived from the choices of the subjects. Whenever a stimulus A was preferred over a stimulus B, A received 1 point, and B -1 point. If A and B were rated equally natural, no points were given. On the average, the Mixdorff/Fujisaki-approach received the best ratings ($\mu = 1.63, \sigma = 0.86$), followed by

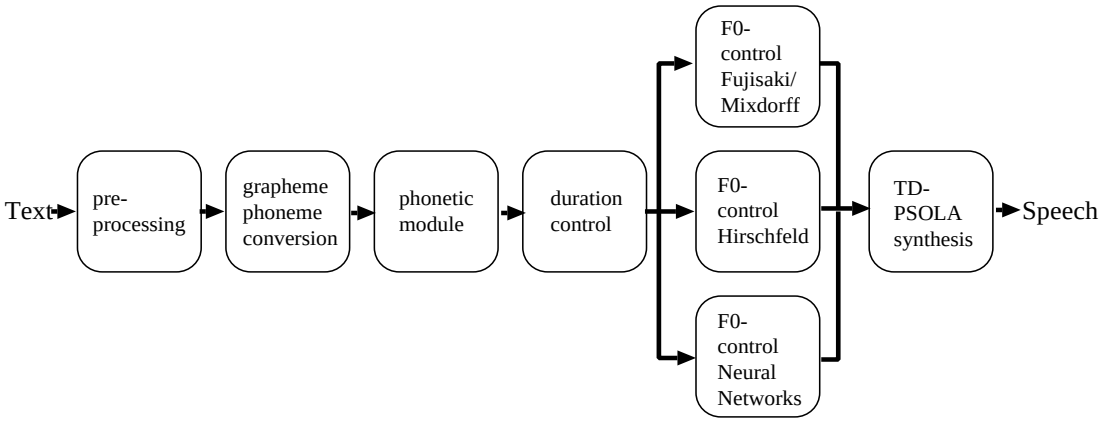


**Fig. 1.** Block diagram of modified TU Dresden TTS-system.

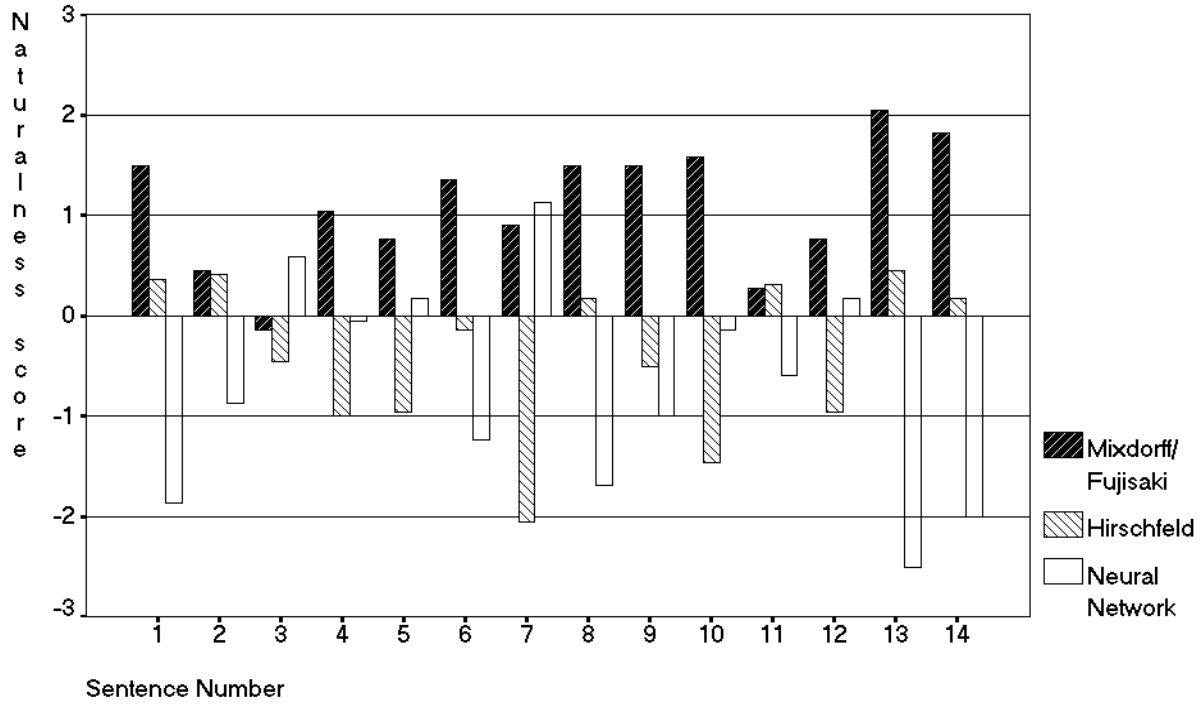


**Fig. 2.** Results of A/B-comparison: Naturalness scores for 14 sentences.

the Hirschfeld-approach ($\mu = -.68, \sigma = 1.12$) and NN($\mu = -.95, \sigma = 1.50$ ). From Figure 2 it can be seen, however, that the ranking varied for individual sentences. For sentences 3 and 7, for instance, NN received the best scoring. It also follows from our results that, since $\sigma$ for the rule-based approaches is much smaller than for the NN, the former produced more consistent results.

## 4. DISCUSSION AND CONCLUSIONS

We presented first results from a comparative study of three different approaches for generating $F_0$ contours. Work is in progress for a detailed evaluation and documentation of all parts of the experiments. Additional experiments for comparing complete TTS-systems concerning their prosodic quality are being prepared.

This project is being sponsored by the Deutsche Forschungsgemeinschaft, reference HO 1674/2-1.

## REFERENCES

[1] Hirschfeld, D. (1996): "The Dresden Text-to-Speech System", 6th Czech-German Workshop on Speech Processing, Prague, Sept. 1996, pp. 22-24.

[2] Mixdorff,H., Fujisaki, H. (1995): A Scheme for a Model-based Synthesis by Rule of F0 Contours of German Utterances. *Proceedings of the '95 Eurospeech*, Madrid, Spanien, Bd. 3, pp. 1823-1826.

[3] Jokisch, O., Pescheck, M. (1997): Neuronale Prosodiegenerierung in der Sprachsynthese, Studientexte zur Sprachkommunikation 14 (Proc. 8. ESSV, Cottbus 1997), pp. 154-161.

[4] Stock, E. et al. (1982): *Deutsche Satzintonation* (VEB Verlag Enzyklopädie, Leipzig).