Exploring the Naturalness of Several German High-Quality-Text-to-Speech Systems

Hansjörg Mixdorff and Dieter Mehnert

Dresden University of Technology
Mommsenstr. 13, 01062 Dresden, Germany

ABSTRACT

The synthesis of near-to-natural F0 contours is an important issue in text-to-speech and crucial to the naturalness and intelligibility of synthetic speech. In earlier studies of the first author a model of German intonation was developed that is based on the quantitative Fujisaki-model. The current paper addresses a perception experiment comparing a TTS-system incorporating this new approach with several German TTS-systems with high segmental quality. Natural speech samples and a synthesis version with natural segment durations were used as references. Results show, that the natural speech samples unanimously received 10 points on a 0 to 10 point scale. The best TTS-systems cluster around a mean value of 5.0, whereas the variant with natural durations reached a mean score of 6.6 points, indicating the importance of closely modeling natural segment durations.

Keywords: TTS, prosody, perceptual evaluation.

1. INTRODUCTION

Generating near-to-natural F0 contours is an important issue in text-to-speech synthesis and crucial to the quality of synthetic speech achieved. In earlier studies by the first author a model of German intonation was developed which is based on the quantitative Fujisaki-model. A typical F0 contour is described as a sequence of major rises and falls, so-called ‘tone-switches’ which are modeled by onsets and offsets of accent commands connected to accented syllables [1, 2]. This model, henceforth referred to as **M**ixdorff-**F**ujisaki-Model of **G**erman **I**ntonation (MFGI), was integrated into the TU Dresden TTS-System (DRESS).

The current paper addresses a perception experiment comparing the prosodic naturalness of DRESS with MFGI, DRESS with a neural network (NN)–based intonation control [3] and three other German TTS-systems with comparably high segmental quality.

Samples of the external systems, henceforth referred to as ‚alien systems A, B and C‘, each of them using a sound inventory produced by a male speaker, were collected from interactive websites. Three of the TTS-systems had PSOLA-, and one LPC-based segmentals. Recorded speech uttered by the speaker who had produced the inventory for DRESS was used as a reference.

An additional variant of DRESS with MFGI with segmental durations copied from the natural utterances was also included in the evaluation.

The experiment follows a series of previous experiments which had been performed within the framework of DRESS [4, 5]

2. DESIGN OF THE EVALUATION

The corpus used for this study was the following news passage consisting of three sentences of a total duration of about 15 seconds produced with each of the systems:

„Hell erleuchtet präsentierte sich am späten Montag abend die Kuppel des Reichstags. Die Bundestags-Baukommission sah sich vor Ort an, wie das künftige Parlamentsgebäude einmal wirken soll.

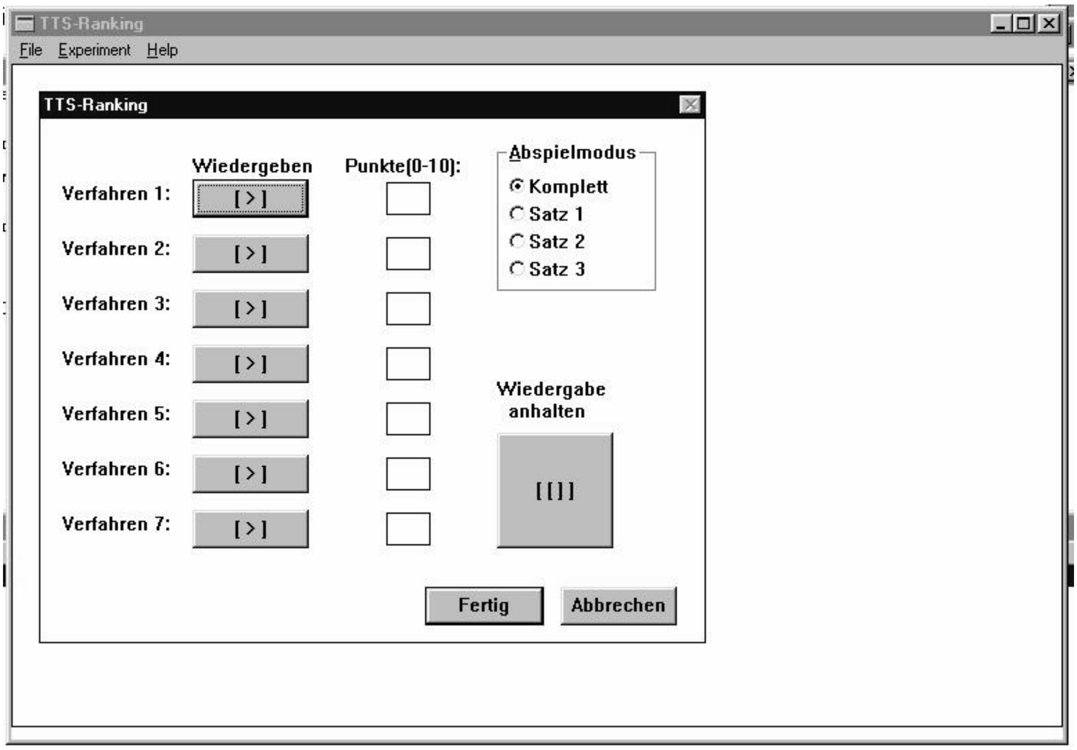


Figure 1. User interface used in the ranking experiment.

Zu besonderen Anlässen wie der Wahl des Bundespräsidenten soll die Kuppel effektvoll leuchten.“ -*“Late Monday evening the cupola of the Reichstag presented itself brightly illuminated. The construction commission of the Bundestag examined in place, what the future parliament building will look like. On special occasions, such as the election of the Federal President, the cupola will be illuminated effectfully.”*

The experiment was conducted using a laptop computer. 24 subjects (13 male, 11 female) with university background took part in the evaluation and were requested to judge the prosodic naturalness of the speech samples. Most of the subjects had already taken part in similar experiments and were familiar with synthetic speech.

After specifying their name, age and sex, the subjects were presented with a GUI equipped with playback buttons for each TTS-system (Figure 1).

Subjects were allowed to play back each of the examples over headphone as often as they liked. To the right of the playback buttons edit fields were located where the subjects were requested to enter the number of points they assigned to the respective systems.

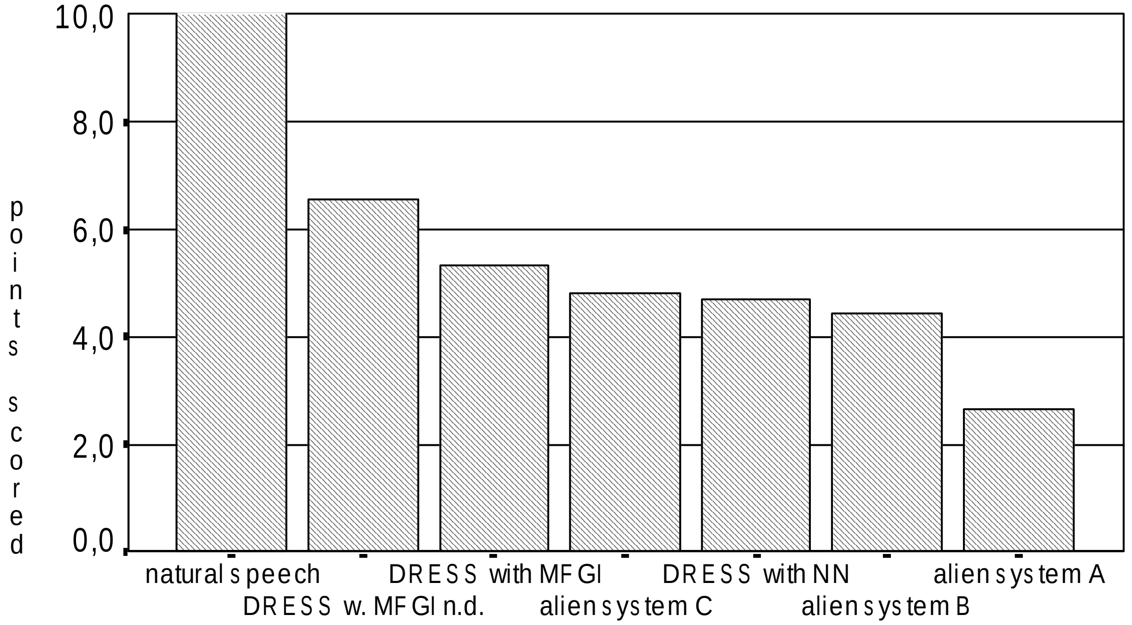
In an additional control field radio buttons were arranged for selecting a playback mode. Subjects could either listen to the whole news passage in one row or to each of the three sentences individually.

They were asked to assign points to each system on a scale between 0 and 10, and were requested to give 0 points to the least acceptable and 10 to the best one. This pre-condition was introduced in order to encourage subjects to make use of the whole range of the scale.

3. RESULTS

As could have been expected, the natural speech samples were rated best and unanimously reached the maximum score of 10.0 (see Figure 2 for mean scores). The best TTS-systems, including DRESS with MFGI cluster around a mean value of 5.0, whereas alien system A falls off considerably. The DRESS/MFGI variant with natural segment durations was preferred to the all-synthesis approaches and reached 6.6 points.

The frequency with which a system was rated least acceptable corresponds to the mean preference score (Table 1).



* DRESS w. MFGI n.d. = DRESS with MFGI and natural durations

Figure 2. Bar plot of mean ratings in the naturalness test.

Table 1. Frequency of judgment ‚0 points'.

System	Frequency „0 points"
natural speech	0
DRESS/MFGI n.d.	1
DRESS/MFGI	2
alien system C	3
DRESS/NN	3
alien system B	4
alien system A	10

Paired-samples T-test was performed and showed that the results stated so far are significant ($p < .05$) or highly significant ($p < .01$). Table 2 gives detailed results.

The total number of points assigned in the experiment varied depending on the subject ($\mu = 38.5$, $\sigma = 5.1$). One subject gave extremely low ratings to all synthetic examples ($\Sigma = 23$).
Many subjects commented that the request to assign 0 points to the least acceptable TTS system posed problems, not necessarily because the distance between systems could not be expressed by the number of points, but because of the negative connotation of ‚0 points', since none of the systems was perceived as being completely inacceptable.
Less skilled subjects observed difficulties in abstracting from the segmental impression and concentrating on the prosodic quality. Hence we cannot exclude that in the case of prosodically equally acceptable examples judgments were influenced by the quality or the pleasantness of the synthetic voice.

4. DISCUSSION AND CONCLUSIONS

First of all it must be stated that the text sample used in this experiment is very short and was selected ad-hoc, and that different material might have led to different results.

The ranking of TTS-systems found in the present experiment differs from the one observed in a previous experiment, a sentence-wise pair-comparison [5], where alien system B had significantly outperformed the other TTS-systems. We had then concluded that this result might be explained by the fact that the F0 range used be system B was wider than that of the other systems

Table 2. Results from paired T-test for significance, h.s. = highly significant, s. = significant, n.s. = not significant.

	DRESS/ MFGI n.d.	DRESS/ MFGI	alien system C	DRESS/ NN	alien system B	alien system A
natural speech	.000 (h.s.)	.000 (h.s.)	.000 (h.s.)	.000 (h.s.)	.000 (h.s.)	.000 (h.s.)
DRESS/ MFGI n.d.		.011 (s.)	.023 (s.)	.005 (h.s.)	.018 (s.)	.000 (h.s.)
DRESS/ MFGI			.514 (n.s.)	.345 (n.s.)	.272 (n.s.)	.009 (h.s.)
alien system C				.878 (n.s.)	.572 (n.s.)	.015 (s.)
DRESS/ NN					.752 (n.s.)	.023 (s.)
alien system B						.039 (s.)

and equalled the range found in natural speech, making system B sound more lively. In the current experiment on a news passage a wide F0 range might have been perceived as overly emotional.

We believe that the design applied in the present study is more realistic than a pair-wise comparison of isolated utterances, since TTS-systems are typically used for reading coherent passages of text, ie news bulletins, traffic information, weather reports, made up of more than a single sentence. If a potential customer wished to select a system for his purposes, he would best be presented with an interactive dialog similar to the one used in this study permitting him to play back the same text passage with each of the TTS-systems available.

We conclude that:

- the synthesis systems DRESS with MFGI and alien systems B and C are perceived as being almost equally natural,
- the quality distance between these systems and natural speech is still great and corresponds to almost half of the scale used.

The fact that the DRESS/MFGI variant with natural durations was found more acceptable than the full synthesis systems indicates that modeling natural segment durations can considerably improve the perceived prosodic quality. This will be the subject of our future research.

This work has been funded by the Deutsche Forschungsgemeinschaft (grant HO 1673/2-1).

5. REFERENCES

[1] 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.

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

[3] Jokisch. O.. Pescheck. M. (1998): Neuronale Prosodiegenerierung - Einfluß der Trainingsdaten. In *Fortschritte der Akustik 1998*. Zürich. Switzerland. pp.352-353.

[4] Mixdorff.H.. Mehnert. D. (1998): Perceptual Evaluation of Three Different Approaches for Generating F0 contours in TTS. In *Fortschritte der Akustik 1998*. Zürich. Switzerland. pp. 398-399.

[5] Mixdorff, H., Mehnert, D. and Hirschfeld, D. (1999): Comparing the Naturalness of Several Approaches for Generating F0 contours in German TTS Systems. In the *Proceedings of the 137th ASA / 25th DAGA-Meeting,* Berlin.