LexDRESS
SPEECH SYNTHESIS FOR A SPEAKING PRONUNCIATION DICTIONARY
FIRST RESULTS

Rüdiger Hoffmann[1], Ursula Hirschfeld[2], Oliver Jokisch[1], Lutz C. Anders[2]

[1] Technische Universität Dresden, Germany
[2] Martin-Luther-Universität Halle – Wittenberg, Germany
{Ruediger.Hoffmann|Oliver.Jokisch}@ias.et.tu-dresden.de
{hirschfeld|anders}@sprechwiss.uni-halle.de

Abstract: The Institute of Speech Sciences and Phonetics at Halle University is developing a new dictionary of German pronunciation which will also incorporate a "talking dictionary". This development is done in co-operation with the Laboratory of Acoustics and Speech Communication at Dresden University of Technology. A preliminary study published at ICPhS 2003 [1] showed that the quality of a standard TTS system (in this case the Dresden Speech Synthesis System DRESS) is not sufficient for the purpose of teaching the standard pronunciation. Therefore, the synthesis system lexDRESS was developed based on DRESS but showing features which are specifically important for a word synthesis from a close phonetic transcription. Using this close transcription requires a speech inventory which considers a large amount of allophonic sounds. A first version of the synthesis system as well as the inventory is available now. It is the aim of this paper to give an overview on lexDRESS. Furthermore, the results of a first evaluation are described here. A set of test words has been evaluated by experts and laypersons with regard to adherence and closeness to norms of pronunciation.

1 Introduction

For almost 50 years the Institute of Speech Sciences and Phonetics at Halle University has been investigating the correct pronunciation of German and its coding. Presently, a team is working on a new dictionary of German pronunciation [2-4]. Part of this project is the development of a "talking dictionary" (CD-ROM), undertaken in co-operation with the Institute of Acoustics and Speech Communication at Dresden University of Technology.

Nowadays, a talking dictionary is a necessary add-on to a conventional dictionary, demanded by users. It makes the acquisition of the standard pronunciation easier since it is difficult for unskilled users to transform phonetic transcription into spoken language. Furthermore, the International Phonetic Transcription, often used in pronunciation dictionaries, is not sufficient for a comprehensive description of all features of standard German pronunciation. For example, the variants of the consonantal /r/ are represented by just one transcription symbol, and with respect to stress only the position of stress is described, not the means by which it is realized.

Recording the entire projected number of sound examples (about 150,000, mainly single words) would be very expensive and consumes as much as 6 Gigabytes of memory space (16 Bit, 16 kHz recording), which is too much even for a standard DVD, let alone a CD-ROM. Under these circumstances, speech synthesis is the solution to be preferred. Even for a

synthesis system with an extended diphone set the expenses are essentially lower (e.g., 5,000 diphones consume approx. 20 Mbytes), and flexibility with respect to further changes and/or additions is guaranteed.

For this purpose, the Dresden Speech Synthesis System DRESS [5, 6] was selected. In a former investigation [1], the suitability of the baseline system was evaluated. Not surprisingly, the results showed that a special version of the system had to be developed to meet all specific requirements of the high-quality synthesis of the talking dictionary.

In the following, we summarize first the disadvantages of standard TTS for our special application. Then, we describe the system lexDRESS. The inventory and the intonation rules are discussed in more detail. Finally, the results of a set of listening experiments which were performed to evaluate the quality of the synthesis are discussed.

2 Disadvantages of Standard TTS

2.1 System Overhead

A standard TTS system like DRESS must be as universal as possible. The preprocessing modules and the grapheme-to-phoneme conversion must be able to process arbitrary text input. Because of the existence of numerous exceptions, however, universal modules are always potential error sources. For the purpose of word synthesis, the structure of the system can be simplified with respect to the following items:

- Because only single words have to be synthesized, the calculation of a phrase or sentence prosody can be omitted.

- Because the phonetic transcription of the word is given in the dictionary, no module for grapheme-phoneme transcription is required.

- The calculation of the word prosody has reliable symbolic input (accent information) from the available transcription.

2.2 Acoustic Quality

In the most cases, diphone based TTS systems are far from reaching high naturalness. This is a special problem for a synthesizer which is to teach standard pronunciation. In our former evaluation [1], we identified a series of disadvantages especially from the phonetic point of view. They can be subdivided into prosodic problems and segmental problems:

- In testing the single words, suprasegmental deviations were generally assessed more negatively than segmental deviations and deviations in consonants were more negatively assessed than deviations or disturbances in vowels. This results were not expected since we assumed that the known problems with prosody manipulation in speech synthesis would not have as much influence on the synthesis of isolated words.

- Deviations from the standard pronunciation occured in synthesized isolated words (with mapped prosody) with respect to several phonetic features. For a complete listing, see [1].

Especially, the definition of the diphone set of our baseline system which is based on a selection of 42 allophones proves to be too coarse.

3 The System LexDRESS

3.1 Software System

As already mentioned, lexDRESS is a derivative of our baseline system DRESS. During the last years, we focused on the development of a version called microDRESS [7] which is especially suited for embedded systems which require small footprints. The code of this system was adapted to meet the requirements of single word synthesis mentioned above.

The concatenation and the prosodic manipulation of the speech units are performed by a PSOLA-like algorithm which was optimized for microDRESS especially [8]. It shows high computational performance combined with good perceptual quality.

3.2 The Inventory

Basing on the experiences with the baseline system, a diphone inventory was defined consisting of the following subsets:

1. VC combinations from 21 vowels (including diphthongs) with 24 consonants (including glottal stop). The diphones were recorded twice (in stressed and unstressed position).

2. CV combinations from 27 consonants (including the glottal stop) with 20 vowels (including diphthongs). The diphones were recorded twice (in stressed and unstressed position).

3. CC combinations from 29 consonants (as far as reasonable).

As an example, the second subset is shown in Table 1 (at the end of this paper). Of course, only reasonable combinations are included. Considering the stressed and unstressed versions, the inventory consists of approx. 2,400 diphones. This is about double the number of diphones in our baseline system.

3.3 Prosody Generation

For a single word synthesis, the prosody model can be simplified essentially. However, models for the prosody of German isolated words play virtually no role in the engineering literature.

For lexDRESS, the following solution was selected as a first step. Each syllable of the word which is to be synthesized is labelled by an accent level, as shown in Table 2. In this way, an accent pattern which can be easily interpreted is produced. For example, the pattern *2124* would be assigned to the word *Mathematik*, etc.

The accent patterns are converted in the prosodic control parameters (pitch, duration, and intensity) according to the internal rules of the acoustic module of DRESS.

Table 1 - This excerpt from the diphone list of the new inventory indicates all diphones from the CV type with an "X". Those diphones which were already included in the inventory of the baseline system are printed in bold. In this way, the extensions made for the new system can be identified. The notation is in X-SAMPA [9].

	i:	**I**	**y:**	**Y**	**e:**	**E:**	**E**	**2:**	**9**	**o:**	**O**	**a:**	**a**	**u:**	**U**	i_^	**@**	**aI**	**aU**	**OY**
P	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
T	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
K	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
B	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
B_0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
D	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
D_0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
G	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
G_0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
?	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**			**X**	**X**	**X**
M	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
N	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
F	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
V	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
V_0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
Ts	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Z	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
Z_0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			X	X	X
S	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
Z	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**			
Z_0	X	X	X	X	X	X	X	X	X	X	X									
C	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**						**X**			
J	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**		**X**	**X**	**X**	**X**
R	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
R_0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
L	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	X	**X**	**X**	**X**	**X**
H	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**			**X**	**X**	**X**

Table 2 - Accent levels used in lexDRESS.

level	meaning
4	(main) accent
3	minor accent
2	no accent
1	no accent and reduced

4 Preliminary Evaluation

4.1 Test Set

We used the same set of test words which was used for the test of the baseline system [1]. This test set was also applied in former investigations and reflects a number of known problems. It is summarized in Table 3.

Table 3 - The 35 words of the test set.

Aids	blühen	Glück	Schularbeiten
Angstzustände	Bronchitis	Häuschen	Sonnenbrand
anzünden	Brötchenkorb	Infekt	spazieren
Ärzte	dahinter	Keuchhusten	Tulpen
Aufbruch	Durchblutungsstörungen	Kinderlähmung	überqueren
Autos	Erkältung	Kopfschmerzen	Verstauchung
Bahnhof	Feldscheune	Magengeschwür	Wadenkrampf
Bandscheibenschaden	Frühlingswetter	Obstbäume	Zahnschmerzen
begeistert	Gehirnerschütterung	Schnitzel	

The words of this test set were synthesized in the following versions:

1. *dress.joerg:* TTS synthesis with the baseline system DRESS using the voice “Joerg”. This is the configuration which was applied for our former investigation [1]. It shall serve for a comparison with the new voice. The prosody is formed by the standard procedures of the baseline system (Klatt durations and Fujisaki intonation).

2. *map.joerg / map.ulrike:* Word synthesis with the new system lexDRESS using the standard voice (Jörg) and the new female voice (Ulrike) with original prosody. In this case, the synthesized speech signal is equipped with a prosodic contour which was copied from the original signal (mapped prosody). It is the purpose of this technique to evaluate the segmental quality separated from the influence of the prosody model.

3. *ldress.joerg / ldress.ulrike:* Word synthesis with the new system lexDRESS using the standard voice (Jörg) and the new female voice (Ulrike) with a synthetic prosodic pattern according to the model described in section 3.3 and in Table 2.

4.2 MOS Test

Because this paper is a workshop report on the very first results with lexDRESS, we are able to present the first evaluation of the quality of the inventory only.

For a statistical evaluation, ten words of the test set described above have been randomly selected and used for a mean opinion score (MOS) test. For the test, the set of words which were synthesized according to the three methods described in section 4.1 was complemented by the same words in natural pronunciation produced by the same female speaker who produced the diphone inventory (*original.ulrike*).

The MOS test was performed by the following listeners:

- 39 laities, 33 male / 6 female, mean age 24.8 (14 … 50)
- 11 experts (in speech synthesis, not in phonetics), 8 male / 3 female, mean age 38.4 (22 … 55)
- total of 50 probands, 41 male / 9 female, mean age 27.8 (14 … 55)

The results of the MOS test are summarized in Figure 1.

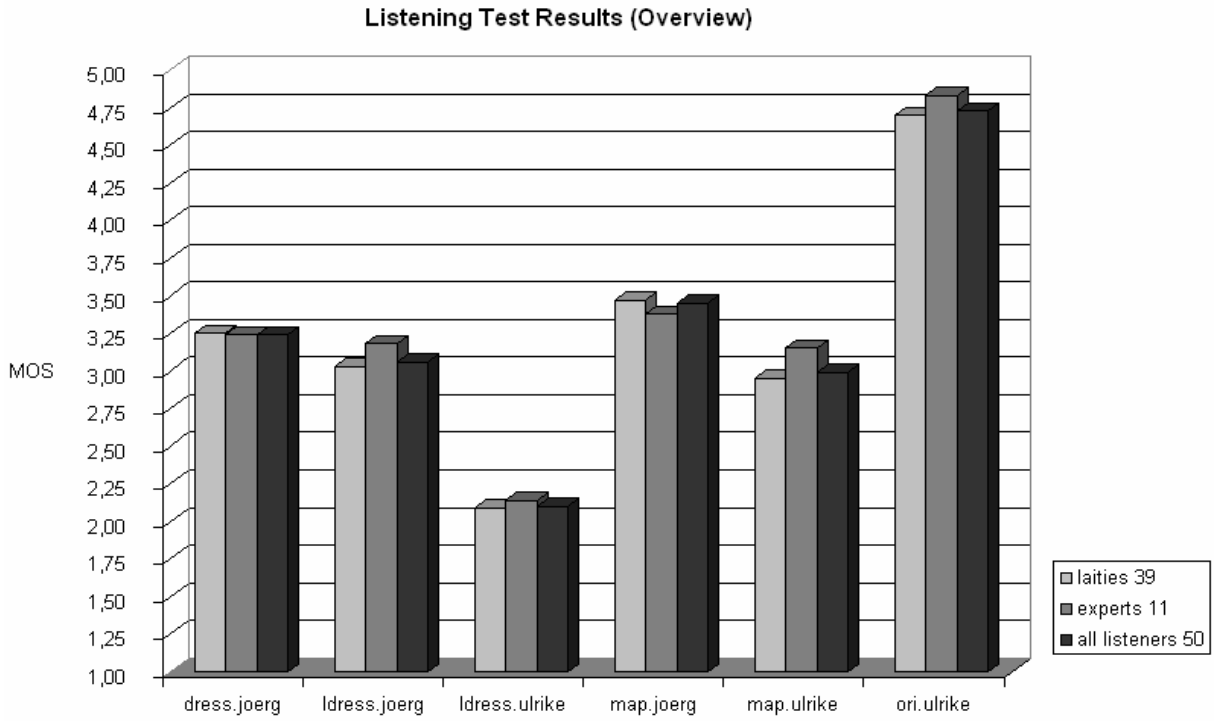

Figure 1 – Results of the preliminary MOS test.

The results can be interpreted as follows:

- Because we used the "rough cut" of the new inventory, the segmental quality is not optimized at all. Obviously, the judgment of the listeners is influenced by such effects.
- The type of the prosodic modeling has a strong influence independent of the voice.

4.3 Judgment by Experts

The material which was used for the MOS test was also evaluated by phoneticians with respect to the intended application (similar to the evaluation of the baseline system [1]). The phoneticians identified a clear progress of the phonetic quality.

Therefore we are confident that the new voice will perform very well after removing the segmental problems of the “rough cut”.

5 Application

LexDRESS is preliminarily applied in a “dictionary demonstrator”. This is an experimental platform which shows in four columns the following information per dictionary entry:

1. the orthographic version of the word,

2. the IPA transcription of the word,

3. an interactive field in which the IPA transcription can be edited or extended by additional information,

4. the X-SAMPA transcription converted from the contents of the previous column applying the rules which were proposed in [9].

Obvioulsly, the columns 1 and 2 form the contents of the printed version of the pronunciation dictionary. Columns 3 and 4 are required for its “speaking” counterpart.

6 Conclusion

It is our aim to improve a diphone based speech synthesis so far that it is suited to demonstrate standard pronunciation. The solution requires special features for single word synthesis on the one hand and a subtle design of the inventory on the other. For further refinement, the following things must be done now:

- According to the results of the first MOS test, the segmental quality of the inventory must be improved.

- The evaluation described in this paper must be continued with a larger listener group with phonetic expertise.

- Depending on the results of these tests, further improvements of the system and/or the inventory will be necessary.

- Probably, some of these improvements will concern features which cannot be expressed by the IPA code. This means the IPA coded transcription will be completed by so-called Extended Information (EI) in the future version.

Furthermore, we hope that the results of the development of a high-quality single word synthesis can improve the performance of complete TTS synthesis.

7 Acknowledgements

The work described here was performed in fruitful cooperation between both of our laboratories. Many people have contributed to the solution. Special thanks to Ulrike Kölsch and Peter Müller (Halle) and Dr. Hongwei Ding, Margitta Lachmann and Daniel Sobe (Dresden).

8 References

[1] Hirschfeld, U., Hoffmann, R., Anders, L. C., Kruschke, H., "Speech Synthesis and Standard Pronunciation of German", Proc. Int. Conf. Phonetic Sciences (ICPhS 03), Barcelona 2003, 2593 – 2596.

[2] Krech, E.-M., "Neukodifizierung der deutschen Standardaussprache. Zur Orthoepie-forschung an der Universität Halle" In: A. Braun, H. R. Masthoff (ed.) Phonetics and its Applications. Wiesbaden 2002, 506 - 515.

[3] Stock, E., Hollmach, U., "Soziophonetische Untersuchungen zur Neukodifikation der deutschen Standardaussprache". In: Norm und Variation, Frankfurt 1997, 105 - 115. (Forum Angewandte Linguistik, Bd. 32).

[4] Krech, E.-M., "Gegenwärtiger Stand und neueste Ergebnisse bei der Erforschung der deutschen Standardaussprache". In: B. J. Kröger u.a. (Hg.) Festschrift für G. Heike. Frankfurt/M. 1998, 227 - 241. (Forum Phoneticum 66)

[5] Hoffmann, R., "A multilingual text-to-speech system", The Phonetician, 80 (1999/II), 5-10.

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

[7] Hoffmann, R., Jokisch, O., Hirschfeld, D., Strecha, G., Kruschke, H., Kordon, U., Koloska, U., "A multilingual TTS system with less than 1 MByte footprint for embedded applications," Proc. IEEE ICASSP, Hong Kong 2003, vol. I, 532 – 535.

[8] Strecha, G., Jokisch, O., Hoffmann, R., "Fast efficient signal manipulation in low-resource TTS synthesis", Proc. Int. Workshop Advances in Speech Technology (AST 2003), Maribor, Slovenia, July 2003.

[9] Wells, J. C., "Computer-coding the IPA: a proposed extension of SAMPA." University College London, http://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf.