PROSODIC MODELING OF SYNTHESISED GERMAN WORDS

Ursula Hirschfeld, Rüdiger Hoffmann‡, Friderike Lange**

* Martin-Luther-Universität Halle-Wittenberg / ‡ Technische Universität Dresden, Germany
e-mail: ursula.hirschfeld@sprechwiss.uni-halle.de / Ruediger.Hoffmann@ias.et.tu-dresden.de / friderike.lange@student.uni-halle.de

ABSTRACT

During the development of an "exemplary" synthesis of words and phonetic words for a "speaking pronunciation dictionary", considerable deviations from German pronunciation norms are being found, particularly in the prosodic field. On the basis of listening experiments new possibilities of modelling accent patterns arranged specifically for the German vocabulary are being tested.

Keywords: prosody, synthesis, accent patterns, norms of pronunciation, phonetic word

1. INTRODUCTION

A project involving the close co-operation of phoneticians and engineering scientists is aimed at developing a phonetically high-quality German speech synthesis system making words and phrases (phonetic words) of a German pronunciation dictionary [1] audible. It is intended simultaneously to establish the prerequisites for the wider use of the speech synthesis system thus produced.

At present no other German "speaking pronunciation dictionary" of this structure and volume (150,000 keywords) exists. No pronouncing dictionary published up to now has any synthesised audio output. The current DUDEN-Rechtschreibung [2], the DUDEN-Fremdwörterbuch [3] as well as various word lists of restricted volume in computer programs for teaching German as a foreign language ([4]) contain only naturally pronounced voice outputs varying markedly in their quality. Bilingual internet dictionaries (cf. [9]) with a voice output based on synthesised language also have unsatisfactory pronunciation quality. They do not justify their claim to present a pronunciation standard worth emulating.

This is therefore the starting point of our investigations and development work.

2. PRELIMINARY WORK

Up to now initial investigations with a high-quality speech synthesis system (Dresden speech synthesis DRESS) have been carried out (cf. [5], [6], [7] [8]), leading to the following conclusions:

- The necessary technology is available (algorithmic component). Since only words and simple phrases are used in the dictionary project, the component necessary for linguistic text analysis can be omitted and a source of unwanted artefacts therefore avoided.
- The voice quality achievable with the presently available configuration level of the system, is, however, still not satisfactory, as shown by several acceptability trials.

Demands for the expansion of the DRESS system within the framework of the project therefore affect the considerations of expert phoneticians during the adaptation of the prosody control of the presently available system, since it is particularly the prosodic control (melody variation, intensity variation, vowel duration) that is not satisfactory.

In traditional text-to-speech (TTS) synthesis, the prosodic parameters are mostly determined using a combination of knowledge-based and neural algorithms. This also applies to DRESS [10]. Although these methods are sufficient to produce intelligible synthetic speech, the quality is not sufficient to demonstrate the standard pronunciation of isolated words. It could be tried to adapt the models to data based on isolated word pronunciation but we did not expect much success due to general drawbacks of automatic learning procedures. We therefore use another approach for the "speaking pronunciation dictionary".

3. CURRENT WORK

At the present time lists of all phonetically relevant accent grading patterns are put together for words and phrases (phonetic words) (cf. 3.1). Rules for the prosodic control of words and phonetic words

are derived in the form of accent templates with variable prosodic parameters. A PC-based experimentation platform referred to as Lex-Editor is used for this (cf. 3.2). The correctness of the linguistic reproduction of the dictionary entries is checked in acceptability trials. (cf. 3.3).

3.1. Accent patterns and grading

The keyword list of a dictionary contains a relatively small variation of accent patterns which result from the syllable number and the accent grading between the syllables of a word and/or a phonetic word. In addition, the morphological-semantic structure plays a major role on prosodic organisation. A distinction has to be made between:

- single words such as *Antwort (answer)*
- determinative compounds such as *Carlsbergstiftung*
- copulative compounds such as *Rheinland-Pfalz (Rhineland-Palatinate)*
- multipartite keywords such as *Santiago de Chile*
- phrases such as *durch dick und dünn gehen (to go through thick and thin with sb).*

The morphological-semantic structure determines the accent patterns (with determinative compounds normally stressed on the first and copulative compounds on the second component), the distribution of primary and secondary accents and the accent grading and organisation. Four grades of accent are applied: 4 – primary accent, 3 – secondary accent, 2 – unstressed, 1 – reduction (unstressed with additional weakening), for example

Gehirnerschütterung (concussion): **1-4-1-3-1-2**

With regard to the accent organisation monosyllabic words are easiest to be handled, where only vowel length needs priority consideration. Disyllabic keywords form the biggest part of the vocabulary. With two strengths of accent and two vowel lengths – short or long accent vowels – there can be these accent patterns:

- short: *kommen (coming),* long: *gehen (going)*
- short: *Bericht (account),* long: *bevor (before).*

In case of three-syllable words the variation possibilities are correspondingly higher. Again, every possibility appears quite frequently:

- short: *antworten (answering),* long: *anbieten, (offering)*
- short: *bekommen (achieving),* long: *gehören, (belonging)*
- short: *Optimist (optimist,* long: *Polizei (police).*

Accent variation increases with the number of syllables but the number of keywords declines sharply.

3.2. Lex-Editor

As already mentioned, the dictionary will be equipped with a special version of the speech synthesis system DRESS which will be called LexDRESS. It generates a speech signal based on the phonetic transcription of the appropriate dictionary entry and the accent pattern provided for it. In the recent phase of optimising the system, an interactive version of LexDRESS, called Lex-Editor, is used. The block diagram of the editor is shown in Figure 1.

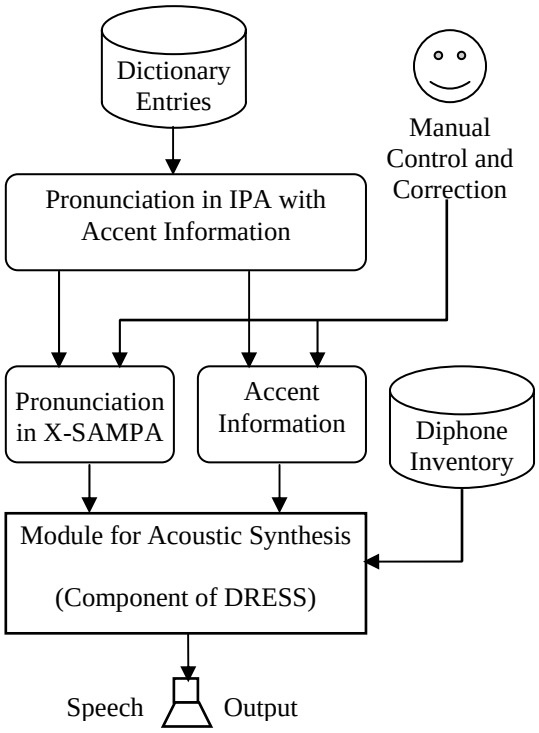
The Lex-Editor consists of the following three main components:

- An interactive component which replaces the linguistic component of the original TTS system DRESS. The module displays the graphemic dictionary entry and the transcript of its pronunciation (including accentuation) in two different codes, IPA and X-SAMPA [11]. The user is able to edit the X-SAMPA sequence and the accentuation scheme of the word in order to correct or to optimise the acoustic output.
- Diphone inventory. A collection of all required sound combinations (diphones) forms the acoustic database of the system. The application of diphones is widespread in speech synthesis because the sound transitions are considered properly in this simple way. The diphones are extracted from words which were naturally pronounced by a female speaker. In general, the number of diphones depends on the extent in which the allophonic variability of the sounds is considered. In the case of LexDRESS, this number is large because the standard pronunciation has to be emulated as close as possible.
- Acoustic synthesis module. This module extracts a sequence of diphones from the diphone inventory according to the X-SAMPA input. These diphones are concatenated to form the desired sound sequence. The pitch, sound

duration and intensity of the sounds are modified according to the accent information which was provided formerly. The synthesis process that transforms the symbolic information to acoustic parameters is completed with this step.

In this way, the material for the following listening tests was prepared.

Figure 1: Structure of the Lex-Editor



3.3. Acceptance investigations

As a basic step for the systematic investigation of accent patterns and grading, acceptance tests were carried out in order to examine the *accent template* used to implement the four accent stages (cf. 3.1). Within the experimentation platform LexDRESS the prosodic organisation of the synthesised words can easily be adjusted by means of the parameters fundamental frequency (F0), loudness (L) and duration (D). Up to now the following programming of the prosodic parameters was used (Accent template 1):

Table 1: Programming of accent template 1

Accent template 1	**F0** (in %)	**L** (in dB)	**D** (in %)
Accent stage 4 (*primary accent*)	20	8	10
Accent stage 3 (*secondary accent*)	10	5	2
Accent stage 2 (*unstressed*)	0	0	0
Accent stage 1 (*unstressed+weakening*)	0	0	-10

In acceptance tests carried out in 2004 and 2005 (cf. [7]) test persons were asked to simply comment on the prosody of several test words synthesised using accent template 1. The results showed considerable deficits in the prosodic organisation of the words as for instance:

- too many secondary accents
- secondary accents too strong
- staccato-like rhythm
- no gradual melody motions, too staggered
- neighbouring syllables too monotonous
- unstressed syllables drawled.

Based on these results accent template 1 was optimised as follows:

- secondary accents too strong → accent stage 3: F0 from 10 to 2; L from 5 to 8
- unstressed syllables appear stretched → accent stage 2: duration from 0 to -10
- melody on unstressed syllables too monotonous → accent stage 1: F0 from 0 to -5

Table 2: Programming of the optimised accent template 2

Accent template 2	**F0** (in %)	**L** (in dB)	**D** (in %)
Accent stage 4 (*primary accent*)	20	8	10
Accent stage 3 (*secondary accent*)	**2**	**8**	2
Accent stage 2 (*unstressed*)	0	0	**-10**
Accent stage 1 (*unstressed+weakening*)	**-5**	0	-10

For the latest acceptance test a series of test words (compounds with secondary stress) was synthesised, each word in three different versions:

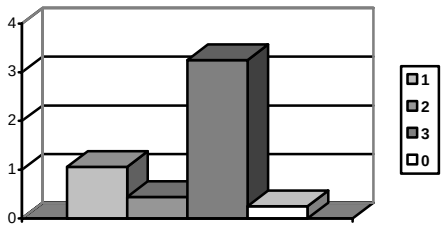
Version 1: Accent template 1
Version 2: Nonsense-accentuation
Version 3: Accent template 2

16 expert listeners were asked to rate their preferences for the different versions of each word. This produced a clear preference for version 3 of all the test words (cf. Figure 2 and Table 3).

Table 3: Preferred versions of test words

Versions of test words	Mean
1 – accent template 1	1,06
2 – nonsense version	0,44
3 – accent template 2	3,25
0 – no version preferred	0,25

Figure 2: Preferred versions of test words



The optimised accent template 2 was clearly preferred even though the differences in fundamental frequency, loudness and duration were only minimal compared to accent template 1.

The same test persons were also asked to comment on anything conspicious concerning the prosodic organisation of the test words synthesised with accent template 2. Still, a large number of notable deficiencies became obvious, which makes further optimisation of the prosodic control unavoidable. This particularly concerned the fine adjustment of the parameters in relation to each other, the smooth tonal transition from syllable to syllable, and the falling tone of words with a stressed final syllable.

4. FURTHER TASKS

At present the following tasks have to be completed:

- systematisation of the vocabulary (150,000 words) into classes of accent patterns and accent grading
- experimental processing of the prosodic features with the Lex-Editor
- development of a rule for the prosodic organisation of the classes of accent patterns and grading
- assigning to each entry of the dictionary an accent pattern as described above
- further optimisation of the diphone inventory especially with regard to the demands made by foreign words that are continuously included into the dictionary.

5. CONCLUSIONS

We described the further development of the idea of a "speaking pronunciation dictionary" using a speech synthesis system which we introduced in [5]. The test and evaluation of an experimentation platform within which prosodic parameters can be modelled with regard to the accent structures are innovative.

6. ACKNOWLEDGMENTS

We would like to thank Oliver Jokisch, Margitta Lachmann and Guntram Strecha (Dresden), who prepared the version of DRESS used in the experiments and where also very much engaged in the realisation of the project.

7. REFERENCES

[1] Krech, E.-M., Stock, E., Anders, L.C., Hirschfeld, U., 2008 (in prep.). *Aussprachwörterbuch* Berlin: de Gruyter.

[2] Duden – *Die deutsche Rechtschreibung* 2006. 24. Auflage. Mannheim u.a.: Dudenverlag.

[3] Duden – *Das Fremdwörterbuch* 2005. Mannheim u.a.: Dudenverlag (CD-ROM).

[4] Esser, O., Klinker, Th. 1996. *Aussprachetraining DaF, CD-ROM für die Grundstufe*, Ismaning: Hueber. / Rausch, R., Rothe, H. 1999. *Besser Deutsch sprechen*. Universität Leipzig. (CD-ROM, Windows-Version)

[5] Hirschfeld, U., Hoffmann, R., Anders, L.C., Kruschke, H. 2003. Speech Synthesis and Standard Pronunciation of German. In: *Proc. 15th ICPhS* Barcelona, 2593-2596.

[6] Hirschfeld, U., Hoffmann, R., Jokisch, O., Lachmann, M. 2005. Speech synthesis for a German pronunciation dictionary – phonetic evaluation. In: Vich, R. (ed.), *Electronic Speech Signal Processing*. TUDpress Dresden, 446-451. (Studientexte zur Sprachkommunikation 36)

[7] Hirschfeld, U., Hoffmann, R. 2006. Standardaussprache per Sprachsynthese? In: Hirschfeld, U., Anders, L.C. (eds.): *Probleme und Perspektiven sprechwissenschaftlicher Arbeit.* Frankfurt/M., 135-146. (Hallesche Schriften zur Sprechwissenschaft und Phonetik 18)

[8] Hoffmann, R., Hirschfeld, U., Jokisch, O., Anders, L.C. 2004. LexDRESS – Speech synthesis for a speaking pronunciation dictionary: First results. In: Fellbaum, K. (ed.), *15. Konferenz Elektronische Sprachsignalverarbeitung*, Cottbus, 183-190. (Studientexte zur Sprachkommunikation 30)

[9] LEO – Online-Wörterbuch Deutsch-Englisch, Deutsch-Französisch, Deutsch - Spanisch: http://www.leo.org/leo_home_de.html visited 29-Dec-06.

[10] Mixdorff, H., Jokisch, O. 2003. Evaluating the quality of an integrated model of German prosody. *International Journal of Speech Technology* 6/ 1, 45-55.

[11] Wells, J.C. 1995. *Computer-coding the IPA: a proposed extension of SAMPA.* University College London, Revised draft, http://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf.