

---

# Automatic Accent and Phrase Boundary Detection for Speech Synthesis Corpora

Diploma Thesis (Informatics) at the University of Hamburg

Timo Baumann <mail@timobaumann.de>

2009-01-13, 10:24

## Abstract of the Diploma Thesis

Prosodic annotation (i.e. marking the speech flow with discrete prosodic symbols) is an important step in the processing of speech data for text-to-speech synthesis. It allows for data-driven prosody modelling in the speech synthesis process which then includes all idiosyncratic prosodic properties of the recorded speaker. Furthermore, modern speech synthesizers can incorporate information about accentuation and phrasing in their unit-selection mechanisms which improves the naturalness of the synthesized speech.

The diploma thesis summarized (originally titled: *Automatische Erkennung von Akzentuierungen und Phrasierungen in Sprachsynthesekorpora*) deals with methods for automatic prosodic annotation using machine learning techniques. Every syllable in the corpus is classified as stressed or unstressed and for every word boundary its status as a phrasal boundary is categorized. Classifiers for these tasks are trained and evaluated on independent test material. More complex prosodic modelling (e.g. the integration of the classifiers' output in context) is left out which allows to test such models against the results of this work.

When looking at prosody, the distinction between phonology and phonetics is kept in mind. Three prosodic systems are distinguished on the phonological level: Accent, Juncture and Intonation. Accent is a lexically determined (binary) property of every syllable. Juncture describes the degree of unity of neighbouring parts of an utterance and helps in the structuring and decoding of the speech flow. Intonation helps in the selection and differentiation of meaning by assigning a speech melody. Intonation selects accents which are then realized in stressed syllables and adds accent tones. Additional phrasal tones support the structuring role of Juncture. The phonological subsystems influence the phonetic phenomena of Accentuation and Phrasing which in turn have effects on the phonetic-prosodic features intensity, melody, duration and pause and their ultimately audible and measureable acoustic correlates in the speech sound. The practical part of the thesis uses acoustical (and other) features in order to determine the status of Accentuation and Phrasing of syllables and word boundaries respectively.

Several prosodically annotated (German) speech synthesis corpora are used as the data base for this work, each containing a (large) number of utterances by one or several speakers. The heterogeneity of the corpora in both the theoretical conception of prosodic annotation, as well as their practical realization (text selection, speaker selection including dialect, automatic or manual segmentation) is shown qualitatively and supported by corresponding measurements. The annotation is then unified as much as possible. All corpora have to be syllabified, i.e. syllable boundaries were automatically deduced from the segmental annotation. Also, the correspondence between written and spoken words has to be reconstructed in order to be able to also use text-based tools for feature extraction.

A large number of features are extracted for Accentuation classification (segmental, syllabic, word-based, duration, pitch and pitch curves) and Phrasing classification (word-based, syntactic, pauses, lengthening, pitch and pitch curves). Feature evaluation uses a wrapper approach around the classification algorithms. Speaker-dependent, speaker-independent and cross-corpus experiments are conducted for Accentuation and Phrasing. Speaker-dependent Accentuation classification proved to be similar for all speakers and corpora (f-measure: 70–80%) and only deteriorated little for speaker-independent (f-measure: 70%) and cross-corpus (f-measure: 65%) experiments. Phrasing classification, on the other hand, does not fair as well: Speaker-dependent results differ strongly and mirror the differences in the annotation of the corpora. Thus, speaker-independent classification leads to weak results. In reverse, it can be concluded that the annotation of Accentuation is sufficiently similar across corpora and can be generated speaker-independently, while the annotation of Phrasing still needs better standardization across corpora to allow for automatic speaker- and corpus-independent annotation.

Finally, a not previously annotated speech synthesis corpus is prosodically annotated and used to train a TTS-system's speaker-dependent prosodic models. A listening experiment shows weak results compared to a baseline system which uses the prosody model of a different speaker. An extensive error analysis shows several problems.

The error-prone annotation of Phrasing, the TTS-system's prosody training (which can only use text-based features while the classifiers find these to be of little use), and a rule-based insertion of pauses optimized for the baseline system's models and very unfitting in the new system.