PROLAB - THE KIEL SYSTEM OF PROSODIC LABELLING

K.J. Kohler
IPDS, Kiel, Germany

ABSTRACT

For the Kiel Corpus of Spontaneous Speech [1] a prosodic labelling system (PROLAB) has been developed. It is based on a prosodic model for German (KIM - The Kiel Intonation Model) and uses a 7 bit ASCII repertoire.

KIM - THE KIEL INTONATION MODEL

The prosody model incorporates the following domains [2,3]:

(1) lexical stress – three levels: unstressed, secondary stress in compounds, and primary stress

(2) sentence stress – four levels: reinforced, accented, partially and completely deaccented

(3) intonation:

- pitch peaks and valleys and their concatenation
- synchronization of pitch peaks and valleys with stressed syllables – three steps: early, medial, late
- downstep of successive pitch peaks/valleys and pitch reset

(4) prosodic boundaries (degrees of cohesion) – three variables: pause duration, phrase-final segmental lengthening, scaling of F0 end points

(5) overall speech rate

(6) disfluencies: pauses, breathing, hesitations

Stress

Within stress we have to differentiate between lexical and sentence stress. At the abstract level of phonological specifications in the lexicon, every German word has at least one vowel that has to be marked as potentially stressable, as being able to attract the feature specifications of sentence stress. Lexical stress is thus not a distinctive stress feature, it only marks a position that can attract such a feature at the sentence level, but need not.

By default, content words are accented and function words completely deaccented. Deviation from default content word stress may be partial or complete deaccentuation, determined by syntax, semantics and pragmatics. Thus, e.g., in 'verb + direct object' constructions the verb is partially deaccented in neutral (non-focussed) accentuation, no matter whether it precedes or follows the object (*Max schreibt einen Brief./Max hat einen Brief geschrieben.*), whereas in 'verb + adverbial' constructions default accentuation is kept (*Max hat täglich geschrieben.*). In either case deviation from this neutral pattern implies focus (of the verb or the adverb, respectively). Complete deaccentuation in the first case introduces focus contrast on the object, which may be strengthened by emphatic reinforcement. Function words, although completely deaccented by default can receive all the content word sentence accent categories by deviation from default.

Intonation

All lexically stressed vowels of words with 'primary' or 'secondary' (= partially deaccented) sentence stress receive intonation features, which may be either 'valleys' or 'peaks', and in the case of 'peaks', they may contain a undirectional F0 fall, or rise again at the end, resulting in a fall-rise. 'Valleys' may have a low rise, to indicate, e.g., continuation, or a high rise, used, e.g., in questions.

All 'peaks' and 'valleys' may have their turning points (F0 maximum in 'peaks', or F0 minimum in 'valleys') early or later with reference to the stressed-vowel onset. For 'peaks' the non-early position may be around the stressed vowel centre (= medial) or towards its end (= late).

Peaks are characterized by a quick F0 rise confined to the vicinity of a sentence-stressed syllable. This rise precedes the onset of the latter, and is usually short and narrow in range, for an early peak; it extends into the first half of the stressed nucleus in the case of a medial peak. In the late peak, it starts after the stressed vowel onset and continues into the second half of the nucleus or beyond; the exact timing of the maximum peak value depends on vowel type (duration according to quantity and quality), subsequent voiced/voiceless consonants and number of immediately following unstressed syllables. There may even be a low stretch of F0 in the stressed vowel before the rise.

Valleys, on the other hand, have a continuous rise, starting before the stressed-syllable nucleus (early) or inside it (non-early) and extending as far as the beginning of the following sentence-stressed syllable. If there are several unstressed syllables between two sentence stresses a valley is thus realised as a more gradual F0 ascent compared with the much quicker rise for a late peak. The less distance there is between stressed syllables the more difficult it becomes to distinguish between a 'valley + peak' and a 'late peak + peak' sequence, especially if there is no F0 dip in between the first and second stress F0 maxima, as in a hat pattern.

In a concatenation of pitch peaks without prosodic boundaries between them, F0 may fall to a low or an intermediate level and then rise again for the next peak. This fall will be effected on intervening unstressed syllables between the two peaks, reaching the lowest point, to start the next rise, in the vicinity of the following stressed syllable, depending on peak position. If there are no unstressed syllables separating the two peaks, the dip can be accommodated between all peak combinations, except for 'late + early/medial', where a hat pattern is created; it combines the rise of the 'late peak' and the fall of the 'early peak' in a two-stress sequence.

This absence of an F0 descent between peaks can also be extended to concatenations with intervening unstressed syllables. In such a hat pattern, an early peak is not possible initially, and a late one is excluded non-initially. If there are more than two stresses incorporated in a hat the non-initial and non-final ones are unspecified as to peak position because they neither have a rise nor a fall but are simply integrated into the downstepped sequence of peak maxima. In the categorization of pitch patterns they are nevertheless grouped together with peaks. If in a two-stress rise-fall it is difficult to decide whether the rise represents a valley, or a late peak in a hat pattern, the latter solution is chosen.

When prosodic boundaries intervene any sequencing of peaks and/or valleys is possible, but the hat pattern is then excluded since it represents a very high degree of cohesion. On the other hand, a late peak with a full F0 descent marks a dissociation from a following peak and will then normally be linked with a prosodic boundary, i.e. final lengthening and F0 reset afterwards.

Unstressed syllables preceding the first sentence stress in a prosodic phrase may be either low or high: they represent different types of pre-head.

Declination, i.e. the temporally fixed decline of F0 has been replaced by downstepping in KIM, i.e. a structurally determined pitch lowering from sentence stress to sentence stress, independent of the time that elapses between them.

Prosodic boundaries

One of the functions of prosody is the sequential structuring of utterances and discourse. Two categories of phrasing have been set up so far. [PG1] corresponding to prosodic sentences and [PG2] related to prosodic phrases. Both are always phonetically signalled by lengthening before them, and usually by F0 resetting after them. Asides and parenthetic insertions have no F0 resetting in spite of other clear phrasing marker signals. Contrariwise, F0 resets may occur at other points than the phrasing markers [PG1,2]. [PG1] also coincides with high syntactic structure nodes, whereas [PG2] does not. Both may be further strengthened by the incidence of pauses and intonation patterns. Full F0 peak descents are particularly frequent with [PG1], and [?] as well as [.?] are only associated with this phrasing marker.

SYMBOLIZATION OF THE MODEL CATEGORIES

The symbolic labelling system has to meet the following requirements:

- unequivocal representation of the categories of the prosodic phonology
- integration into 7 bit ASCII segmental label files
- integration into 7 bit ASCII orthographic files of German text
- clear typographic separation from the segmental labelling allowing prosodic notations on the same tier for convenient cross-reference between segmental and prosodic aspects of speech
- mnemonic ease for learning and use.

The application of these guiding principles has resulted in the standardization of the following repertoire and conventions [4] for insertion in orthographic text or segmental phonetic files.

- Apostrophe and quotation mark ['], ["] for lexical stress are put in front of the primary or secondary stress vowel; unmarked vowels are unstressed. In a segmental label file these stress markers are linked to the vowel symbol, in an orthographic file they are inserted in logical order before, and on the same time mark as, the vowel. Function words, marked by suffixed [+], do not get a lexical stress symbol by default; if they receive sentence stress, double apostrophe [' '] is inserted before the vowel of the appropriate syllable.
- Digits [3],[2],[1],[0], when not combined with punctuation marks, refer to sentence stress. They are put in logical order before words that receive the reinforced, accented, partially or completely deaccented sentence stress category. The lexical stress position then determines where F0 contours have to be hooked.
- Punctuation marks [.],[,],[?] refer to pitch peaks, low and high rising valleys, and the character sequences [.,] and [.?] to the corresponding fall-rises. They are put in logical order before a prosodic boundary or before the next sentence-stress digit [>1]. [(.)?] can only occur before a prosodic boundary.
- Parentheses [)],[(] refer to early and late peaks or early and non-early valleys and are put after the sentence-stress digit; the medial peak is marked by the absence of these symbols. Digit and parenthesis form a symbolic unit.
- The pitch movement between successive peaks or between a peak and a boundary may be a full or an intermediate F0 descent or a level F0, symbolized by digits [2],[1],[0] before [.]. Digit and punctuation mark form a symbolic unit.
- Downstep is not marked. F0 reset is implied by a prosodic boundary; in the case of its absence, [=] is prefixed to the next digit [=>2]. If reset occurs at other points than boundaries, [+] is prefixed to the next stress digit [=>2].
- A high prehead is marked by [HP] at the beginning of an utterance or after a phrase boundary.
- Prosodic phrasing markers [PG1] and [PG2] are put after punctuation marks at the appropriate places.
- Only speech rate changes in relation to the speed in the preceding prosodic phrasing unit are marked: [RP] and [RM] (= 'rate plus/minus') are put after [PG1,2] (and before [HP]). An absolute rate judgement at the utterance onset may be added at a later labelling stage.
- Disfluency markers are

    -- [z:] for hesitation lengthening at the end or inside of a word
    -- [/+] or [=/+] for break-offs and resumptions at word boundaries and within words, respectively.
- Markers for segmental phrase-level units are [p:], [h:] (= pause, breathing), [l:], [s:] etc. (= laughing, clicks etc.) [4].
- All non-segmental prosodic markers are without duration; they are put on the same time mark as the beginning of the next segmental unit.

LABELLING PROCEDURE

A labelling platform has been created at IPDS by M. Pätzold on an AT, running on UNIX and equipped with a sound card, which accepts segmental label files, generated by the KTH MIX programme, and F0 analysis data as input, allows the display of F0 contours and labels, as well as the insertion, deletion and change of prosodic labels under auditory and visual control. The default sentence stress markers [2] for content words and [0] for function words and a general prosodic phrasing marker [PG] are inserted automatically on the basis of the segmental labels. The manual labelling then proceeds in cycles dealing with one prosodic domain after another. The result is a label file that integrates prosodic labels into the segmental strings. The following orthographic transcript with prosodic annotations (rather than a complete label file, to reduce the amount of information and for greater ease of intelligibility) provides an illustration of the prosodic labelling of a spontaneous dialogue from the Kiel Corpus of Spontaneous Speech [1].

g071a004.s1h
TIS004:
2 <{hm> **PG2 2(** D'ienstag **0** w}rde+
0 mir+ **0** g'ut **1. 2)** p'assen **2. PG1**
2 <{hm> **PG2 0** das+ **2** h'ei~t **, PG2 p:**
2 Mom'ent **1. PG2 2** 'allerdings **,**
2 'erst **z: , PG2 2(** n'achm"ittags **h: 2.**
PG1 RP HP 0 das+ **0** wird+ **0** dann+
2 wahrsch'einlich **0** 'n+ **0** b'i~chen **0.**
2) schw'ierig **2. PG1 2** D'ienstag **1.**
2 m'ittwochs **z: 1. PG2 RM** <{h> **PG2**
p: 0 is=/+ **0** s'ieht **0** das+ **0** bei+
2 m' 'ir+ **z: 0** sch=/+ **2. 2** schw'ierig
0 'aus **2. PG1 RP 0** da+ **0** hab' **0** ich+
=**2** tags'}ber **1. 2** Term'ine **1. PG1 RM**
h: 2 <{hm> **PG2 HP 0** wie+ **0** s'ieht
0 das+ **0** bei+ **2** ' 'Ihnen+ **0** am+ **1.**
3 D'onnerstag **0** 'aus **2. PG1**

Prosodic label files can now be the input to the RULSYS/INFOVOX TTS system for German, which also contains an implementation of KIM [3], to test the adequacy of the manual labelling by comparing its rule synthesis with the original. Prosodic modelling, prosodic labelling and prosodic synthesis thus form an integrated framework of prosodic research at IPDS Kiel. The prosodic categories, being related to human sound production beyond the particular language phenomena found in German, are transferable to the description of other languages, and PROLAB may be used more generally in prosodic labelling.

ACKNOWLEDGEMENT

Part of the work reported here was carried out with financial support from the German Ministry of Education, Science, Research and Technology (BMBF) under VERBMOBIL contract 01IV101M7.

REFERENCES

[1] IPDS (1995), *CD-ROM#2: The Kiel Corpus of Spontaneous Speech*, vol. I, Kiel: IPDS.

[2] Kohler, K.J. (1991): "A model of German intonation", *AIPUK*, vol. 25, pp. 295-360.

[3] Kohler, K.J. (forthcoming): "Parametric control of prosodic variables by symbolic input in TTS synthesis", 2nd *ESCA IEEE Workshop on Speech Synthesis*, Sept 1994, New Paltz, N. Y.

[4] Kohler, K.J., Pätzold, M., Simpson, A. (1994), *Handbuch zur Segmentation und Etikettierung von Spontansprache - 2.3.* VERBMOBIL Technisches Dokument Nr. 16, Kiel: IPDS.