Modelling Prosody in Spontaneous Speech

Klaus J. Kohler
IPDS Kiel
University of Kiel, Germany

ABSTRACT Following on from general considerations of requirements for prosodic modelling of spontaneous speech, this paper outlines a prosody model for German, its incorporation in a TTS system as a prosody research tool, the model-based development of a prosodic labelling system for the application to spontaneous speech, and the use of the resulting prosodic label files as input to the TTS system for transcription verification and model elaboration.

1 Introduction

Whereas speech analysis and modelling have traditionally focussed on scripted speech (logatomes and words in isolation or in standard sentence frames, connected speech in sentences and texts), spontaneous speech is now receiving increased attention, also in the area of prosody. But - at least initially - work in this new field of research proceeds on the assumption that the theoretical categories and operations on them, established for scripted speech, can simply be transferred to spontaneous speech. This will most certainly require adjusting in two ways: some categories will no longer be adequate (deletion operating over time being a case in point), and new categories and operations will have to be added (e.g. in connection with dysfluencies and repairs).

Furthermore, the modelling of prosody will have to take the following points into account.

(1) Prosodic universals
The study of prosody has grown out of dealing with individual languages, especially with English, more than with any other language. Categories and operations (e.g. prosodic rules) are - to a large extent - determined by the particular linguistic structures. What we need for a general prosodic theory, however, are independently motivated categories and operations.
Candidates are pitch direction (falling, rising) and synchronization of pitch 'peaks' and 'valleys' with syllable timing, in each case independently of the functional use they may be put to in individual languages (e.g. tone or intonation), further, the prominence-lending features of pitch movement and segment duration for the functional use of sentence stress (focus), and timing factors at various levels (global speech rate, utterance-final lengthening, stress/syllable timing).

(2) A unified theory integrating segmental and prosodic aspects, as well as intonation and timing among the latter.

We have become used to thinking in and dealing with dichotomies in the study of speech: segmentals versus prosody, intonation versus timing. It is quite clear that all these levels of description form an intricate mutually conditioning network. Prosody (e.g. stress, timing, especially speech rate) provides conditioning factors for articulatory reduction; on the other hand segmental structures determine the manifestation of prosodic categories (different synchronization of pitch 'peaks' and 'valleys' in high/low or short/long vowels, curtailing of falls in pitch peaks before voiceless consonants). Global utterance timing has not only an influence on individual segment durations, but also on their qualitative realization and on the manifestation of pitch patterns (upward scaling of F0 and reduction of F0 range in increased speed). Contrariwise, segmental features determine their timing in global utterance speed: vowel and consonant durations are adjusted differently, not by a uniform proportionate factor across all segments; depending on the segmental type and context, shortening in fast speech also involves assimilation and elision of articulatory gestures over and above their changes in timing.

(3) A prosodic phonology as an interlevel between syntax/semantics/pragmatics and the phonetic signal: phonetic substance - phonological form - linguistic function.

The phonetic-semantic relationship is not direct in the sense that the measured values themselves represent syntactic or semantic categories, but that the link operates via formal elements that, on the one hand, are related to features of meaning, but are, on the other hand, defined by phonetic ranges. This phonetic substantiation of phonological categories is just as essential as the recognition of structure in phonetic substance. Both phonetic substance and phonetic structure (or signal measures and phonological form) are required for an adequate description of the phonetic-syntactic/semantic relationship, and consequently for prosodic modelling.

(4) Integration of prosodic modelling into a linguistic environment.

It follows from (3) that prosodic modelling requires strong links with syntactic, semantic, and, particularly in the case of spontaneous speech, also
pragmatic levels.

This contribution outlines
- a prosody model that has been developed for German in accordance
  with the four requirements specified for prosodic modelling: KIM - the
  Kiel Intonation Model
- its incorporation in a TTS system (RULSYS/INFOVOX) as a prosody
  research tool
- the model-based development of a prosodic labelling system (PRO-
  LAB) to create prosodic label files of spontaneous speech, which may
  in turn be used as input into the TTS system for transcription verifica-
  tion and model elaboration.

2 A prosodic phonology of German: the Kiel
Intonation Model (KIM)

2.0 The categories of the model and its general structure

This chapter gives a summary of Kohler (1991a,b, 1995), supplemented by
the most recent additions that became necessary to cope with spontaneous
speech phenomena encountered in prosodic labelling. The following domains
have to be recognised in a prosodic model of German:
(1) lexical stress
(2) sentence stress
(3) intonation:
   (a) categories of pitch 'peaks' and 'valleys' as well as their combina-
       tions at each sentence-stress position
   (b) types of pitch category concatenation
   (c) pitch of pre-head preceding the first sentence stress in a prosodic
       phrase
(4) synchronization of pitch 'peaks' and 'valleys' with stressed syllables
(5) 'downstep' of successive pitch 'peaks'/'valleys' and pitch reset
(6) 'upstep' of successive pitch 'peaks'/'valleys'
(7) prosodic boundaries (degrees of cohesion)
(8) overall speech rate (changes) between the utterance beginning and suc-
    cessive prosodic boundaries
(9) register change
(10) dysfluencies: pauses, breathing, hesitations, break-offs and resump-
    tions.

A system of prosodic distinctive features is used to specify the abstract
phonological categories in these domains, and they enter into sets of ordered
symbolic rules. The features are either graded or binary and determine the
parametric value spaces activated by parametric phonetic rules following the
symbolic ones. The prosodic features are attributed to phonological units,
which are either segmental (vowels and consonants) or non-segmental (morphological and phrase boundaries). Attached to vowels is the fundamental distinction within the German prosodic system, viz. stress and intonation. In separating symbolic rules from subsequent parametric ones for generating acoustic output, KIM recognizes two levels in prosodic modeling:
- the defining of phonology-controlled prosodic patterns by a small number of significant F0 points (macroprosody),
- the output of continuous F0 contours influenced by articulation-related modifications (microprosody, for further details see Kohler 1995).

KIM is integrated into a pragmatic, semantic and syntactic environment. The input into the model are symbolic strings in phonetic notation with additional pragmatic, semantic and syntactic markers. The pragmatic and semantic markers trigger, e.g., the pragmatically or semantically conditioned use of 'peak' and 'valley' types or of sentence focus. Lexical stress position can largely be derived by rule, and syntactic structure rules mark deaccentuation and emphasis in word, phrase, clause and sentence construction. Phrasal accentuations are thus derived from the syntactic component preceding the prosodic model, and are given special symbolizations in the input strings to the model (see Kohler 1991a,b for further details).

2.1 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. This lexical stress is thus not a distinctive stress feature, it only marks a position that can attract such a feature at the sentence level. In non-compounds as well as in at least one compound element of compounds there is one vowel with such 'primary' lexical stress; other compound elements have one vowel each with 'secondary' lexical stress. All remaining vowels are lexically 'unstressed'.

Sentence stress is attributed to a word as a whole, and manifests itself phonetically at the lexical stress position. By default a non-function word receives the category 'accented'. Deviations from this are either in the direction of emphatic 'reinforcement', or of 'deaccentuation', which may be 'partial' or 'complete'. Function words are by default 'unaccented' (= 'completely deaccented'). Deviations are 'partially (de)accented', 'accented' or 'reinforced'.

Vowels receive combinations of the stress features +/-FSTRESS> and +/-DSTRESS> (referring to the association of sentence stress with the two important parameter domains of F0 and duration). In sentence-stressed words, the vowel with 'primary' lexical stress is +/-FSTRESS, +DSTRESS>; in 'completely deaccented' content words it is +/-FSTRESS, -DSTRESS>. Vowels with 'secondary' lexical stress are also +/-FSTRESS, +DSTRESS>, irrespective of sentence stress.
Finally, in 'unaccented' function words as well as in lexically 'unstressed' syllables, the combination is \(<\text{FSTRESS},\text{DSTRESS}\)> . In 'partially deaccented' sentence stresses \(<\text{DSTRESS},\text{DEACC}\)> is added to the two positive stress features, all other vowels are \(<\text{DEACC}\) . Words that are to get additional emphasis receive the feature \(<\text{EMPH}\) in their lexical stress position, all other vowels \(<\text{EMPH}\). Whether \(<\text{DSTRESS}\), responsible for longer duration, is associated with \(<\text{FSTRESS}\), marking the vowel as the recipient of intonation features ('peak' and 'valley' contours), or as \(<\text{FSTRESS}\), not providing the vowel with this potential, depends on the rules of grammar and context of situation in speech communication, which allocate sentence stress digit markings in the input string to the prosodic model. They have to be supplied by the linguistic environment of the prosodic phonology (see Kohler 1994a, b). The same applies to the attribution of \(<\text{DEACC}\).

To distinguish degrees of emphasis, \(<\text{EMPH}\) vowels may be given the graded stress level feature \(<\@\text{STRLEV}\), with \(@ @ = 1, 2, ... 7\); \(<\text{EMPH}\) vowels are \(<\@\text{STRLEV}\). These vowels are made the more prominent, the higher the stress level. In 'peak' contours, this greater prominence is achieved by raising the F0 maximum, and, if the 'peak' is non-final in a 'peak' series, by having a faster descent as well as by lowering the F0 minimum between 'peaks', proportionally to stress level. In the case of F0 'valley' contours, the final F0 point is raised in accordance with stress level. Emphasis is used to put words and phrases within sentences in focus, particularly when the expansion of intonation contours on certain structural elements is coupled with the deaccentuation of others. \(<\text{EMPH}\) and \(<\@\text{STRLEV}\) associated with \(<\text{FSTRESS}\) do not automatically change the duration linked to \(<\text{DSTRESS}\) . The parametric variation of \(<\text{DSTRESS}\) may be controlled independently of the other stress features; in the model this is captured by the categories of speech rate and hesitation lengthening (2.4, 2.6).

In summary, the following distinctive sentence-stress features are proposed for a comprehensive contrastive categorization in the prosodic phonology of German:
\(<\text{FSTRESS}\>
\(<\text{DSTRESS}\>
\(<\text{DEACC}\>
\(<\text{EMPH}\>
\(<\@\text{STRLEV}\), with \(@ = 0, 1, ... 7\).

The feature pair \(<\text{EMPH}\) and the graded feature \(<\@\text{STRLEV}\) constitute the link with the intonation features. The following tree graph represents the hierarchical relationship between the various sentence-stress features.
2.2 Intonation

2.2.1 Pitch categories at sentence stresses

All vowels with 'accented' or 'reinforced' or 'partially deaccented' sentence stress, i.e. with the feature specification <+FSTRESS,+DSTRESS> receive intonation features, which may be either 'valleys' or 'peaks', specified as <+/-VALLEY>, and in the case of 'peaks' (<-VALLEY>), they may contain a unidirectional F0 fall, classified as <+TERMIN>, or rise again at the end, resulting in a (rise-) fall-rise, categorized as <-TERMIN>. <+VALLEY> is <-TERMIN> by definition. <-TERMIN> may have a low, narrow rise, to indicate, e.g., continuation, or a high, wide rise, used, e.g., in questions, with the specifications <+/-QUEST>. 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 onset of <+VOK,+FSTRESS>, categorized as <+/-EARLY>, and finally, for 'peaks' <+/-EARLY> may be around the stressed vowel centre or towards its end, classified by the feature opposition <+/-LATE>. The categorization of <+VALLEY> into <+EARLY> and <+EARLY>, with a further subdivision of the latter into <+/-LATE>, captures the grouping of 'late' and 'medial' vs. 'early peaks', as it showed up in perceptual experiments with stepwise 'peak' shift from left to right (Kohler 1990b).

Thus the distinctive intonation features needed in the prosodic phonology of German are:

<+/-VALLEY>
<+/-TERMIN>
<+/-QUEST>
<+/-EARLY>
<+/-LATE>
<+/-EMPH>
<+@STRLEV>, with @ = 0, 1,...7.
Their hierarchical relationships, linked to sentence stress, are represented by the following tree graph.

\[
\begin{align*}
&+\text{FSTRESS}, +\text{DISTRESS}, -\text{EMPH}, -\text{DEACC} \\
&+\text{FSTRESS}, +\text{DISTRESS}, -\text{EMPH}, +\text{DEACC} \\
&+\text{FSTRESS}, +\text{DISTRESS}, +\text{EMPH}, -\text{DEACC}, @\text{STRLEV} \\
&-\text{VALLEY} \\
&+\text{VALLEY} \\
&+\text{TERMIN} \\
&-\text{TERMIN} \\
&+\text{QUEST} \\
&-\text{QUEST} \\
&\text{only phrase-final} \\
&+\text{EARLY} \\
&-\text{EARLY} \\
&+\text{LATE} \\
&-\text{LATE}
\end{align*}
\]

The resulting feature combinations for the intonation categories are illustrated in Figure 3:

(a1) \(<-\text{VALLEY}, +\text{TERMIN}, +\text{EARLY}>\): 'early peak'

(a2,3) \(<-\text{VALLEY}, +\text{TERMIN}, -\text{EARLY}, -/+\text{LATE}>\): 'medial/late peak'

(b1,2) \(<-\text{VALLEY}, -\text{TERMIN}, -/+\text{QUEST}, +\text{EARLY}>\): 'low/high non-terminal early peak'

(b3,4,5,6) \(<-\text{VALLEY}, -\text{TERMIN}, -/+\text{QUEST}, -\text{EARLY}, -/+\text{LATE}>\): 'low/high non-terminal medial/late peak'

(c1,2,3,4) \(<+\text{VALLEY}, -\text{TERMIN}, -/+\text{QUEST}, +/-\text{EARLY}>\): 'early/non-early low/high valley'

Which feature combinations are to be activated in the prosodic model again depends on the rules of grammar and context of situation in speech communication, which allocate the intonation markings in the input string. They have to be supplied by the linguistic environment of the prosodic phonology (see Kohler 1991a,b).

2.2.2 Pitch category concatenation and pre-head

In a concatenation of pitch 'peaks' without prosodic boundaries between them (see 2.3), F0 may fall to a low or an intermediate level and then rise again for the next 'peak'. There is also the boundary case of the absence of an F0 descent between 'peaks', which results in a 'hat pattern' ('t Hart et al. 1990). The slight fall due to 'downstep' between the 'peaks' (see 2.2.4) justifies subsuming this boundary case under the intonation feature \(<+\text{TERMIN}>\). In such an intonation structure an 'early peak' is not
possible initially, and a 'late' one is excluded non-initially. The boundary case of zero peak-descent is also applied to a high level F0 before a phrase boundary (see Figure 3 (d1)). In \(<<\text{TERMIN}>>\) 'peaks' the same differentiation between full and intermediate F0 descents must be made (see Figures 3 (d2,3)).

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 are different types of pre-head.

2.2.3 Temporal alignment of 'peaks' and 'valleys'

Taking the default, 'medial peak' as a reference, two significant F0 points are defined. The first one, TF0, is positioned at the beginning of the syllable containing the \(<+\text{FSTRESS}>\) vowel, the second, T2F0, near the vowel centre, the exact timing after voiced vowel onset depending on vowel quantity, vowel height, number of following unstressed syllables and position in the utterance. The calculation of the time point T2F0 after vowel onset is carried out on the basis of the segmental duration rules for German. They have adopted the principle proposed by Klatt (1979) for the rule synthesis of English (see also Kohler 1988), defining different classes of segments (e.g. diphthongs vs. long vs. short vowels, low vs. high vowels) by different pairs of values for intrinsic duration (Di) and for minimal duration (Dmin) and generating actual segment durations in various segmental, prosodic and syntactic contexts by the application of the following rules:

\[
\begin{align*}
(1) \text{<DUR>} \cdot \text{<(Di-Dmin)*PRCNT/100+Dmin>} \\
(2) \text{<PRCNT>} \cdot \text{<PRCNT*PRCNT1/100>}
\end{align*}
\]

In (1), PRCNT = 100 initially; the rules then change the PRCNT values successively by introducing a rule-specific PRCNT1 value into (2). This way all the factors influencing segmental durations (tempo, position in the word and sentence, stress, segmental context) can be captured in specific rules by inserting a new PRCNT1 value each time. This model assumes that all the factors affecting duration operate independently of each other and that it is only the amount exceeding the minimal duration of a segment that is adjusted by these factors. The two assumptions provide a good approximation of segment timing in languages like German and English, and certainly result in prosodically acceptable speech synthesis.

T2F0 for 'medial peaks' is now derived from the basic vowel-type related duration. The only percentage factor that enters the calculation is the one referring to speech rate; it is normally set at 100, a speeding up lowers, a slowing down increases the factor, i.e. it is essentially the intrinsic vowel
duration that determines the point in time after \(<+VOK,+FSTRESS>\) onset (T2F0) where the 'medial peak' is positioned. But this has to be adjusted in the case of aspiration. On the one hand aspiration lengthens the total vowel duration, compared with vowels in non-aspirated contexts, but this increase is not as large as the total aspiration phase; on the other hand it shortens the stop closure duration compared with unaspirated cases, but again not by the total amount. So the larger part of the aspiration (AH) should be added to the vowel, but some of it attached to the plosive, and the F0 'peak' placement has to take this ambivalence into account:

\( \langle +VOK,+FSTRESS \rangle \sim \langle T2F0=((Di-Dmin)^{PRCNT}/100+Dmin)^{0.6+TLAH}^{0.75}\rangle \)

i.e. three quarters of the period up to the last aspiration time point are added to T2F0, shifting it further to the right by this amount.

Sentence-final 'medial peaks' receive a third F0 point, T3F0, at 150 ms after the 'peak' maximum in a medium speech rate (see 2.4); in all non-final cases, the default treatment of a descending F0 is that the 'peak' summit of one \(<+FSTRESS>\) connects with the left-base point of the next \(<+FSTRESS>\). Other possibilities are that the low F0 point between 'peaks' occurs at any other intervening word, or that there is a relatively level "dip" in between.

As the absolute F0 'peak' position is not affected by vowel duration modifications due to voiced/voiceless context, number of syllables in the word, sentence position etc., its relative position changes with vowel shortening or lengthening, moving closer towards or further away from, the end. This way the microprosodic F0 truncation before voiceless obstruents is automatically built into the rules. That no longer applies to rising F0. The intended high value of a 'valley' always has to be physically reached.

An 'early peak' has its maximum value at the \(<+FSTRESS>\) syllable onset, T\(F0\) 100 ms before, and T3F0 - in sentence-final position - in an area where the 'medial peak' has its maximum. A 'late peak' has T\(F0\) at the same point as a 'medial peak', then an additional low F0 point T2F0 is inserted at vowel onset, and the late summit (T3F0) occurs 100 ms after the point where a 'medial peak' has its centre, or at the end of the last voiced segment in a non-final monosyllabic word if this distance is less than 100 ms. If there is an unstressed syllable following, the summit coincides with the unstressed vowel voice onset. In utterance-final position, a fourth F0 point, T4F0, occurs 100 ms after the summit. In monosyllables without final voiced consonants, T3F0 has to occur at least 30 ms before vowel offset to signal the F0 descent to T4F0.

'Valleys' have their left and centre F0 points at the same positions as T\(F0\) and T2F0 in 'medial peaks' (except for 'early low valleys' where T\(F0\) as in 'early peak'); in an 'early valley' the left point is the lowest, whereas in a 'non-early valley' it is the centre point. In both cases the right high point is located at the end of the last voiced segment.
2.2.4 ‘Downstep’ and pitch reset

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. The downstepping values used in KIM are 6% from ‘peak’ to ‘peak’ (starting from 130 Hz in a male voice), and 18% from a ‘peak’ to the next base. In ‘valleys’ both the low and the high F0 value are downstepped by 6%. Downstepping can be interrupted at any point by the feature <+EMPH>, or by resetting. Prosodic boundaries (see 2.3.3) are usually associated with pitch reset but they need not be. Reset can also occur at other points than phrase boundaries.

2.2.5 ‘Upstep’

Besides the interruption of automatic ‘downstep’ at any point by a controlled restart of the downstepping pattern (pitch reset), we also have to take another systematic deviation from default into account, namely the step-wise upward trend of ‘peak’ or ‘valley’ sequences: ‘upstep’. It is treated as a global superpositional feature in KIM and in its TTS implementation (see 3.1). The upstepping values used in KIM and in its TTS implementation are comparable to the ones for downstepping: 6% up from ‘peak’ to ‘peak’, and 12% down from a ‘peak’ to the next base. In ‘valleys’ both the low and the high F0 value are upstepped by 6%.

2.3 Prosodic boundaries

One of the functions of prosody is the sequential structuring of utterances and discourse, i.e. the signalling of prosodic boundaries and - at least partially - their hierarchical organization. To decode the syntagmatic chunking of messages in accordance with the speaker’s intention the listener requires signals that index degrees of cohesion or separation, respectively, between phrases, clauses, utterances and turns. The parameters that achieve this are pause duration, phrase-final segmental lengthening and scaling of F0 and the pause boundaries at the respective boundaries. They can be controlled by parameteric rules in the prosodic model upon appropriate symbolic input.

As at this stage the linguistically and phonetically relevant categorization of these boundaries is not well understood the modelling cannot reduce the categories in this domain to the same small number as in the other areas of prosody discussed so far, but has to allow sufficient degrees of freedom for experimentation with data modelling in a development system, which will be discussed in 3. Two categories of phrasing have been extracted so far with this very flexible device: [PG1] corresponding to prosodic clauses and [PG2] related to prosodic phrases. Both are always phonetically signalled by lengthening before them, and usually by F0 resetting after them. 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 <-TERMIN, +QUEST> is only associated with this phrasing marker.

2.4 Speech rate

The modelling of different absolute speech rates - 'slow', 'medium', 'fast' - within the same speaker, discussed in Kohler (1995), also includes articulatory reduction and elaboration, and recognises a fourth degree, viz. 'reduced' at an otherwise medium rate (see also Kohler (1990s)). This level will probably have to be subdivided into subcategories according to the degree of formality and spontaneity of speaking, comprising different rule modules for the respective degrees. A good deal more research into spontaneous speech is necessary before an adequate categorization can be set up in the model.

This modelling of speech rate and degrees of reduction is extremely useful at the development level, using the TTS research tool (see 3.). As a basis for the description of spontaneous speech data and their labelling, however, it is more helpful to set up relative categories of speech rate change, indicating slowing down or speeding up with regard to a preceding stretch of speech (see 4.). Eventually the two approaches should be combined, for instance in such a way that for a particular speaker's speech production (e.g. a dialogue turn) an initial absolute rate evaluation is provided, upon which slowing down and speeding up operate iteratively.

2.5 Register change

A prosodic model also has to incorporate the category of register (change), in the sense that a speaker may observe equivalent contrastive phonological pitch relations on different pitch levels, with semantic and pragmatic implications, e.g. a lowering of pitch level for asides, and insertions into main arguments, or a raising of pitch level for putting whole stretches of utterance into perspective. It is an open research question at this stage as to how many register levels are needed in the modelling of spontaneous speech. The research tool of 3. uses three: default - raised - lowered. The values used for register in KIM and in its TTS implementation (see 3.) are 20% up for the raised level and 20% down for the lowered one, compared with the F0 points in default register.

2.6 Dysfluencies

At the segmental level, pauses, breathing, hesitation particles, laughing, clicks etc. (see Kohler et al. 1995) need to be indicated as elements of utterance structuring and dysfluency. At the prosodic level, hesitation lengthen-
ing is to be differentiated from automatic phrase-final lengthening. Break-
offs with and without repairs, inside words and at word boundaries are
additional dysfluency categories, characteristic of spontaneous speech, with
the potential of phonetic exponents (see Kohler et al. 1995).

3 A TTS implementation of the model as a prosody
research tool

The incorporation of KIM in the RULSYS/INFOVOX TTS system for Ger-
man has been described in more detail in Kohler (1996) (see also Carlson
et al. 1990). This environment constitutes a powerful research tool in the
analysis and modelling of German prosody. This particularly applies to the
investigation into prosodic phrasing (2.3), amount of F0 descent in 'peaks',
especially in 'peak' concatenation (2.2.2), and into 'downstep and 'pitch
reset' (2.2.4).

To test these aspects the development system uses a prosodic boundary
(cohesion) marker [p:], which is put after the word at which boundary indi-
ces occur. It is preceded by two digits, ranging from [0] to [2], the second
of which refers to pause length, the first to utterance-final lengthening ([0]
standing for absence in both cases). In the case of pitch 'peaks', there is a
third boundary-related digit to the left of these two, referring to the scaling
of the F0 end point, again ranging from [0] to [2]: [2] refers to a descent
to the bottom of the speaker's voice range, [1] to an intermediate fall and
[0] to the absence of an F0 dip, e.g. between the 'peaks' of a 'hat pattern'.
This digit string is preceded by a further digit, ranging from [0] to [3] to
mark four degrees of speech rate from fast to slow. The 'high pre-head'
index is also linked to the prosodic boundary marker in TTS: [p:=]. As our
knowledge of prosodic boundary marking increases the degrees of freedom
can be reduced by establishing constraints between the parameters in the
signalling of the necessary and sufficient number of phonologically relevant
distinctions. The need to retain a high degree of flexibility in the prosodic
development system is also the reason why, in the implementation of KIM
in TTS synthesis, prosodic boundary markings are always preceded by [+],
if there is pitch reset, or they remain unspecified, if there is not.

'Upstep' and three levels of register have also been incorporated in KIM-
based TTS, for a simulation of natural, especially spontaneous speech, but
apart from pauses and hesitation lengthening, dysfluency indices have not
been successfully implemented yet. The extension of 'upstep' is marked by
indexing the two sentence-stressed words where it begins and ends. The
devices used are [8+] and [+8] after the sentence-stress digit [2].

Figures 1-5 of the Appendix illustrate the TTS parameters F0 and dur-
tion for the various KIM-based prosodic categories in this development sys-
tem (RULSYS). They provide F0 (in Hz) for the significant points (square
parameter) as well as cosine interpolation, and the TTS phonetic transcription aligned to the time scale (segment durations in s).

4 The analysis of spontaneous speech

4.1 PROLAB - a KIM-based labelling system

The categories of KIM outlined in 2. have also been the basis for the development of a symbolic system for consistent, systematic and efficient prosodic labelling of recorded spontaneous speech data: PROLAB. This system meets the following requirements:

- unequivocal representation of the categories of the prosodic phonology
- integration into 7 bit ASCII segmental label files
- integration into 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 suitability for easy learning and use
- easy retrieval of prosodic phenomena in data bank searches.

The application of these guiding principles to prosodic labelling of data banks has resulted in the standardization of the following repertoire and conventions (Kohler et al. 1995) for insertion in orthographic text or segmental phonetic files.

- [''],[] 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 sequential order before, and on the same time mark as, the vowel. Function words, identified with suffixed ['+'], by default do not get a lexical stress symbol; if they receive sentence stress, [''] is inserted before the vowel of the appropriate syllable.

- All sentence-prosodic markers are preceded by & to separate them unequivocally from non-prosodic labels, e.g. grammatically determined punctuation marks. The latter are taken over from the orthography and kept as such beside the inserted and &-prefixed punctuations, which refer to intonation categories. Prosodic punctuations follow orthographic ones in their sequential ordering.

- Digits [1,3],[&1],[&0] when not combined with punctuation marks, refer to sentence stress. 'They are put in sequential 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 sequential order before a prosodic
boundary or before the next sentence-stress label [&<digit> ≥ 1].
[&.] and [&.,] occur before the next word after the F0 maximum, i.e.
they are also possible before the sentence-stress label [&<0]. [&(.)?] can
only occur before a prosodic boundary.
- Parentheses [)] and [() refer to 'early' and 'late peaks', the corresponding
  brackets [] and [()] to 'early' and 'non-early valleys'; they are put after
  the sentence-stress digit, e.g. [&2]]; the 'medial peak' is also posi-
  tively marked by ['], which differs from the default implication in 2.2.1
  and in the TTS implementation. It allows easier access in data bank
  retrieval. The same applies to the differentiation between parentheses
  and brackets for 'peaks' and 'valleys'. Sentence-stress digit and pitch
  synchronization marker form a prosodic label 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 before [.] [&<2], [&<1], [&<0]. Digit
  and punctuation mark form a prosodic label unit.
- A high phrase is marked by [&IP] at the beginning of a prosodic
  phrase, but in sequential order after a phrasing marker.
- 'Downstep' is not marked. F0 reset is implied by a prosodic boundary;
  in the case of its absence, [=] is prefixed to the phrasing marker. If reset
  occurs at other points than boundaries, [+] is prefixed to the sentence-
  stress digit, where the reset occurs. In both cases the character forms
  a label unit with the prosodic symbol it is prefixed to.
- 'Upstep' is marked by [1], prefixed to each relevant sentence-stress digit,
  with which it forms a prosodic label unit.
- Prosodic phrasing markers [&PG1] and [&PG2] are put after punc-
  tuation marks at the appropriate places. A phrasing marker that is
  associated with break-offs and resumptions, [-] or [+], is indexed as
  [&PG/]. Asides and insertions into main clauses are indicated by
  bracketed [&PG1][-][&PG1>].
- Only speech rate changes in relation to the speed in the preceding
  prosodic phrasing unit are marked: [RP] and [RM] (= 'rate plus'/
  'rate minus') are put after [PG1/2] (and before [IP]). An absolute rate
  judgement at the utterance onset may be added at a later labelling
  stage.
- Register is not marked yet in PROLAB.
- Dysfluency markers are
  - [2:] for hesitation lengthenings at the end or inside of a word
  - [/-] or [=/-] for break-offs, and [/+] or [=/+] for break-offs and
    resumptions at word boundaries and within words, respectively.
- Markers for segmental phrase-level units are [p], [h] (= pause, breath-
  ing), [l], [s] etc. (= laughing, clicks etc.) (see Kohler et al 1995).
- All non-segmental prosodic markers are without duration; they are put
  on the same time mark as the beginning of the next segmental unit
  (usually word beginning) or as the end of the last segmental unit in the
speech file.

A labelling platform has been created at IPDS by M. Pätzold on an AT, running under X11 and equipped with a sound card, which accepts segmental and/or orthographic label files and F0 analysis data as input, allows the display of the speech wave form, of F0 contours and labels (see Figure 6), as well as the insertion, deletion and change of prosodic labels under auditory and visual control. The manual labelling proceeds in cycles dealing with one prosodic domain after another, in the progression from phrasing to sentence stress, to intonation patterns (peaks, valleys), to their alignment, to speech rate changes and to the other variables listed above. The output is a label file that integrates prosodic labels into the segmental and/or orthographic strings. In this cyclical progression from broad to narrow the labelling can stop at any degree of delicacy, defined by the purpose the resulting label files are to be put to. PROLAB is a comprehensive labelling frame for all prosodic variables, but it is at the same time very flexible to allow a wide range of detail and complexity for different applications.

This PROLAB labelling platform, which also includes a formal check program for the correct syntax of labels in label files, has been applied to a large data base of spontaneous dialogues - The Kiel Corpus of Spontaneous Speech (IPDS 1995). The 31 dialogues on CD-ROM#2 have been provided with prosodic label files in addition to the segmental label files on the CD itself. 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 turn from this corpus.

g071a004    TIS004:
&2 <ähm> &PG2 &2 ( D ’ienstag &0 würdle+ &0 mir+ &0 g ’ut &1. &2) p ’assen &2. &PG1 &2 <ähm> &PG2 &0 das+ &2) h ’ieft &. &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’inlich &0 ’n+ &0 b ’ichen &0. &2) schw ’ierig &2. &PG1 &2) D ’ienstag &1. &2) m ’ittwochs z: &1. &PG2 &2) <ähm> &PG2 p: &0 is=/+ &PG/ &0 s ’ieht &0 das+ &0 bei+ &2) ’m ’ir+ z: &0 sch=/+ &0. &PG/ &2) schw ’ierig &0 ’aus &2. &PG1 &0 da+ &0 hab’ &0 ich+ &2) ’tags ’über &1. &2) ’Term ’ine &1. &PG1 &RM h: &2 <ähm> &PG2 &HP &0 wie+ &0 s ’ieht &0 das+ &0 bei+ &2) ’Thnen+ &0 am+ &1. &3) D ’onnerstag &0 ’aus &2. &PG1

Figure 6 of the Appendix gives the labelled speech waves and F0 contours of the first and the last prosodic clause of this dialogue turn as examples of the PROLAB platform output, demonstrating the application of sentence stress and intonation categories to spontaneous speech.

¹The CD-ROM as well as the prosodic label files may be obtained from IPDS Kiel.
4.2 Transcription verification and model elaboration

Prosodic label files can now be the input to the RULSYS/INFOVOX TTS system for German (see 3.) to test the adequacy of the manual labelling by comparing its rule synthesis with the original. Not all aspects of spontaneous speech in the area of dysfluencies have been successfully implemented in TTS. Work on them is in progress. But even so the synthesis of spontaneous speech from PROLAB files already produces very convincing, natural sounding results that can be evaluated against the original human production. This comparative assessment also provides feedback for improved prosodic modelling of spontaneous speech. Figure 7 of the Appendix provides TTS parameters and speech wave output, as well as its F0 analysis for the same PROLAB representations as in Figure 6. In TTS, the microprosodic F0 is largely controlled by a separate parameter, not shown in the F0 displays of significant points.

Prosodic modelling, its TTS implementation for model testing, prosodic labelling on the basis of the model, prosodic resynthesis of these prosodic label files for transcription verification and renewed model testing and elaboration thus form an integrated framework of prosodic research at IPDS Kiel. The prosodic categories, being related to human sound production beyond particular language phenomena found in German, should also be transferable to the description of other languages, and the portable PROLAB platform be of more general interest in the prosodic labelling of a wide variety of language data.

5 Acknowledgements

This chapter is a revised and expanded version of a plenary paper "Modelling intonation, timing and segmental reduction in spontaneous speech" which I presented at the ATR International Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing in April 1995. My special thanks are due to the organizers for their kind invitation and generous support. Part of the spontaneous data recording and labelling was carried out with funding from the German Ministry of Education, Science, Research and Technology (BMBF) under VERBMOBIL contract 01IV101M7.
6 References


Appendix

Figure 1
Lexical stress and compounding

Figure 2
Sentence stress
Modelling Prosody in Spontaneous Speech

(a1,2,3) 'early/medial/late peak':
#) #ja. ja. #(#ja.

(b1,3,4) 'low non-terminal early/medial/late peak':
#) #ja, ja., #(#ja.,
(b5) 'high non-terminal medial peak': ja.?

(c1,2,3,4) 'early/non-early low/high valley': #) # ja,
#(# ja, #)# ja?
#(# ja?

(d1) 'medial peak with zero descent': ja 0.
(d2,3) 'full/intermediate descent in
<~TERMIN> peak':
ja 2., ja 1.,

Figure 3
Intonation (cf. 2.2.1)
"10 - 2 x 3":
#2# zehn #110p:# minus #2# zwei #000p:# mal #2# drei.
dipped F0 pattern between 'peaks' followed by 'hat pattern'

"(10 - 2) x 3":
#2# zehn #000p:# minus #2# zwei #110p:# mal #2# drei.
'hat pattern' followed by dipped F0 pattern between 'peaks'

Figure 4
Prosodic phrase boundaries
Modelling Prosody in Spontaneous Speech

Downstep default: rote gelbe bleue schwarze #212p: #.

Pitch reset: rote gelbe #+110p: # blau bleue schwarze #212p: #.

Upstep: #2$+# rote gelbe bleue #2+$# schwarze #212p: #.

Figure 5
Downstep, reset, upstep
Figure 6
Labelled speech wave and F0 contour of the first and the last prosodic clause in dialogue turn g071a004
Figure 7
TTS parameters (F0, duration), speech wave output, and its F0 analysis for the PROLAB input of Figure 6