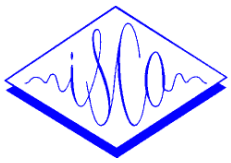

Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

PARAMETRIC CONTROL OF PROSODIC VARIABLES BY SYMBOLIC INPUT IN TTS SYNTHESIS

K. J. Kohler, IPDS Kiel, Germany

Abstract: The Kiel prosodic model for German (KIM), its TTS implementation and the development of a research tool for prosodic modelling and synthesis are outlined.

1. Categories of a prosodic model for German and their symbolization

A prosodic model for German has been developed at IPDS Kiel (= KIM - the Kiel Intonation Model). It takes the following domains into account:

(1) **lexical stress** - two levels: primary, and secondary in compounds
(2) **sentence stress** - four levels: reinforced, neutral, partially and completely deaccented
(3) **intonation:**
    (a) categories of pitch peaks and valleys as well as their combinations at each sentence stress position
    (b) types of pitch category concatenation
(4) **synchronization of pitch peaks and valleys** with stressed syllables - three steps: early, medial, late
(5) **prosodic phrase boundaries (degrees of cohesion)** - three variables: pause duration, phrase-final segmental lengthening, scaling of F0 end points
(6) **overall speech rate** between the utterance beginning and successive prosodic boundaries - four degrees: slow, medium, reduced, fast.
(7) **downstep of successive pitch peaks/valleys and pitch reset**

The following 7bit ASCII characters are used to represent these categories.

(1′) Apostrophe and quotation mark ['"] are put in front of the primary or secondary stress vowel:
R'ück#s"icht ("view to the back") vs. *R'ücksicht* ("consideration")
([#] marks the phonetically - especially prosodically - relevant word boundary in compounds).

(2′) Digits [3,2,1,0] are put in front of words that receive the reinforced, neutral, partially or completely deaccented sentence stress category, which in turn affects the manifestation of the respective lexically stressed vowel. Function words, marked by suffixed [+], have [0] as their default, non-function words [2]; in both cases the digit may be omitted in the symbolization:
2Max Ohat+ Oeinen+ 2Brief 2geschrieben .
("Max did write a letter")
2Max Ohat+ Oeinen+ 2Brief 1geschrieben .
(semantically unmarked rendering of "Max wrote a letter")
2Max Ohat+ Oeinen+ 2Brief Ogeschrieben .
("Max wrote a letter, not a card")
2Max Ohat+ Oeinen+ 3Brief Ogeschrieben .
(reinforcement of contrasted "Brief" in the previous example)

(3′) Punctuation marks [.,?] for pitch peaks, low and high rising valleys, and the character sequences [.,] and [.?] for fall-rises are put

before phrase boundaries (see (5')):
Ja .p: Ja ,p: Ja ?p: Ja .,p: Ja .?p:

(4') Parentheses [)(] for early and late peak positions in sentence-stress syllables are put before the stressed word (after the sentence-stress digit); the medial peak position is regarded as the default case and remains unmarked:
Sie+ hat+ ja+ 2)gelogen .
("She's been lying." = summarizing, concluding statement)
Sie+ hat+ ja+ 2gelogen .
(=start of a new argumentation)
Sie+ hat+ ja+ 2(gelogen .
(as the preceding example but with a contradictory note)

In connection with valleys there are only early and non-early positions in the sentence-stressed syllable. Either one or the other category may be taken as the unmarked default case, depending on their frequency of occurrence: early for [,], late for [?].

(5') The prosodic phrase boundary (cohesion) marker [p:] is put after the word at which it occurs. It is preceded by two digits, the second of which refers to pause length, the first to utterance-final lengthening. 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. Each of the digits may range from [0] (=absence of pause, of final lengthening or of F0 descent) through [1] (= short pause - 200ms; default utterance-final lengthening; intermediate F0 descent) to [2] (= long pause - 500ms; hesitation lengthening; full F0 descent):
zehn .0-210-2p: minus zwei .0-200p: mal drei ("10 - 2 x 3")
zehn .0-200p: minus zwei .0-210-2p: mal drei ("(10 - 2) x 3")

(6') The digit string associated with the phrase boundary marker [p:] is preceded by a further digit, ranging from 0 to 3 to mark four degrees of speech rate, which include degrees of reduction or elaboration : [2] refers to medium overall speed and default reduction, [1] to the same speed but a higher degree of reduction; for [0], degrees of reduction and speed are increased, for [3] they are both decreased from [2]. In this modelling of speech rate, segment durations are not changed by uniform and proportionate up or down-scaling across the whole sequence, but vowels and consonants are dealt with separately according to sets of rules including segmental reduction, assimilation and elision.
mit roten gelben blauen schwarzen .0-3212p:
("with red, yellow, blue and black ones")

(7') The model does not include the category of declination over time, but incorporates the structurally determined, time-independent category of downstep from peak to peak and from valley to valley. It is set at a constant value (6% in medium and slow speeds, 4% in fast speed) and is not indicated symbolically. Pitch reset can occur at any point in the chain of peaks or valleys and is associated with a prosodic boundary. It is marked by [=] before the digit sequence at the preceding [p:].
mit roten gelben .=2110p: blauen schwarzen .2212p:

Since this prosodic symbolization system is based on a prosodic model for

German, it can be used for consistent, systematic and efficient prosodic labelling of recorded speech data. For further details on KIM and its symbolization system see Kohler 1991, 1992, 1994.

2. Implementation of the model in the RULSYS/INFOVOX TTS system for German

KIM has been implemented in the RULSYS/INFOVOX TTS for German. The Kiel development of this TTS system (for details see Carlson et al. 1990, Kohler 1991) makes use of a very simple adaptation of 7bit ASCII to the phonetic transcription of German:
(a) upper-case letters for segmental phonemes, (b) lower case ones for allophones, (c) the characters listed in 1. (1')-(7'). These phonetic symbols are either derived by rule from orthographic input, or they are entered into the system directly, enclosed between the metacharacter [#]. In the latter case the input string can be either entirely phonetic, or mixed orthographic/phonetic as illustrated by the examples in 1.

The greater part of the prosodic notations in (2')-(7') have to be entered as such, because the syntactic component of the system is not powerful enough to derive them by rule from orthographic input. Moreover, in many cases semantic and pragmatic rules would be required to generate the correct prosodic output. The symbolic prosody markers trigger hierarchical sets of parametric F0 and duration rules in the phonetic-to-acoustic output component. In the case of F0, the rules define significant points for peak and valley configurations, synchronize them with lexically stress-marked vowels according to sentence-stress and intonation symbolizations, and modify them contextually as well as microprosodically. A cosine function then interpolates between the final sequence of significant F0 values.

The speed control digit at the [p:] marking attributes a parametric rate variable to every segmental symbol and sets it to a value representing the respective category. Blocks of duration, segment and F0 rules in the phonetic module are then activated by the particular rate variable value and the appropriate calculations along the three phonetic scales are performed. This means that for a particular speed it is not only the segment durations that are adjusted across the whole chain to which the particular rate factor applies, but F0 is also raised for speeding up or lowered for slowing down, and segmental reductions or elaborations are effected simultaneously, in accordance with natural speech production. The segment durations are scaled separately for vowels and consonants and also as a function of a number of other conditioning factors (vowel height, consonant category, stress, number of syllables in the word). The digit before [p:] controlling phrase-final lengthening triggers a more local increase or decrease of segment durations within the set global speech rate.

The TTS implementation of KIM thus allows the modelling of speech timing at a hierarchy of levels from segment to segment chain to phrase to utterance, and this can be achieved entirely by symbolic input referring to the relevant categories of the model, in the same way as this is done for the modelling of stress and intonation. The calculation of individual segment durations follows the Klatt model for segment timing with the formula

$$\text{<DUR>} \;\hat{}\; \text{<(Di-Dmin)} * f_1 * f_2 \ldots + \text{Dmin>}$$

where Di refers to the intrinsic, Dmin to the minimal duration of the segment. f_1, f_2, ... are factors determined by stress, utterance position,

number of syllables in the word and overall speech rate.

The Kiel prosodic model for German is comprehensive and detailed enough for its TTS realization to be capable of generating highly intelligible and natural sounding synthetic output for very intricate phrasing structures in complicated continuous text.

3. A development system for prosodic modelling and synthesis

The TTS implementation of KIM constitutes a research tool for the further development of the prosodic model as well as for the improvement of prosodic synthesis. The categories of the prosodic phonology of German, based on extensive speech production and perception experiments, can be tested in quick interactive auditory evaluations and in formal, more costly listening experiments with carefully prepared synthetic speech output files. The development system allows rule-driven parameter control by symbolic input categories of the model as well as systematic changes of values in graphic parameter displays, in both cases with immediate acoustic output. It is thus possible to check (a) the validity of category differentiations, (b) the need for category extensions or reductions, (c) the adequacy of defined parameter values for the categories. The results can be incorporated in a revised version of the prosodic TTS rules, and this loop of interactive or formal listening test evaluations and rule adjustments can be repeated until an optimization in the intelligibility and naturalness of the synthetic output is achieved.

By referring parametric variables to phonological categories of a prosodic model the degrees of freedom of the TTS generation have been reduced considerably, without foregoing the flexibility and potentially exhaustive coverage of empirical speech data, which a generative framework provides. However, the degrees of freedom in category combination and concatenation, e.g. in connection with phrase boundaries, are still quite large. The TTS development system offers a powerful device for the testing of constraints between prosodic categories. The application of the TTS research platform to prosodic model and synthesis evaluation is also guided by natural speech data labelled within the same category and symbolization framework. So prosodic modelling, its TTS implementation and testing as well as model-driven labelling of natural speech data form an interrelated and mutually conditioning set of procedures in prosodic research at IPDS Kiel.

4. References

Carlson, R., Granström, B. Hunnicutt, S. (1990): Multilingual text-to-speech development and application. In: W.A. Ainsworth (ed.), *Advances in Speech, Hearing and Language Processing*. London: JAI Press, pp. 269-296.

Kohler, K.J. (1991): A model of German Intonation. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)* 25, pp. 295-360.

Kohler, K.J. (1992): Prosodisches Transkriptionssystem für die Etikettierung von Sprachsignaldaten. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)* 26, pp. 238-250.

Kohler, K.J. (1994): Lexica of the Kiel PHONDAT Corpus: Read Speech. Vols. I, II. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)* 27, 28.