SYNTHESIZING PROSODY: A PROMINENCE-BASED APPROACH

Barbara Heuft, Thomas Portele***

*now: Lernout & Hauspie Speech Products, Ieper, Belgium, email: barbara.heuft@lhs.be
**Institut für Kommunikationsforschung und Phonetik, Universität Bonn, Germany

ABSTRACT

A method for generating acoustic prosody is presented that starts from a very simple symbolic input. We present evidence that prominence is a central factor influencing both perception and acoustic parameters. Results of statistical analysis of a large speech corpus are shown, these results have led to the development of a rule system that predicts fundamental frequency and syllable duration. Besides the prominence of syllables and boundaries, position, context and syllable structure are considered by these rules. Finally, the outcome of two evaluation experiments is presented.

1. MOTIVATION

We assume that the prosody of any utterance can be described by assigning only two features to syllables and boundaries: content (phonemes for syllables; rising/falling for boundaries) and prominence (each syllable and each word boundary needs a prominence value). Thus, prominence is regarded as an intermedate parameter between linguistics and acoustics. It is a gradual parameter, although it's optimal range is not yet clear. Beyond the perceptive motivation, this approach has several advantages concerning applications: It is, for example, ideal for synthesizing focal structures (needed e.g. in dialogue systems) or to higher-level text organization (paragraphs of different levels). The approach is described in more detail in Portele & Heuft (1996).

The main demand on the prosody component of such a synthesis system is to convert this simple input into rules for the generation of the acoustic prosodic parameters.

2. PERCEPTIVE PROMINENCE

Several perception experiments have been carried out to validate this aproach by determining: a) the discriminative ability of listeners in the range of our definition and, b) the prosodic correlates of perceived prominence. An experiment concerning the prominence of syllables has been presented in Portele & Heuft (1995); subjects had to rate the prominence of syllables on a scale from 0 to 31. Strong (rho > 0.8) correlations between subjects proved their discriminative ability; and a clear, almost linear relationship between prosodic parameters like syllable duration, height and position of F_0 peaks and perceived prominence of syllables could be shown (see section 4). Comparable results are reported by Fant & Kruckenberg (1989).

A parallel experiment to determine the perceptive prominence of prosodic boundaries has been carried out (Heuft et al., 1996). Subjects had to rate the perceived prominence of the boundary after each word on a scale from 0 to 9. Again, we found very strong correlations between the ratings. The ability to rate consistently perceived prominence of prosodic boundaries is also reported in De Pijper & Sanderman, 1994). The prominence of boundaries was cued by different degrees of final lengthening and the duration of pauses (rho > 0.8). F_0 cues seem to be less important for the perceptive prominence of boundaries (see section 4).

3. IMPLEMENTATION

3.1 The model

Each F_0 contour is regarded as a sequence of peaks. Each peak is described by four parameters: the distance of the peak in ms from the onset of the stressed vowel (*peak delay*); the *height* of the peak [a value between 0 and 1, where 0 is the speaker's F_0 *baseline* and 1 is the *topline*) are the highest and lowest frequency of a given speaker]; and the *slopes* preceding and following the peaks. Minima are described implicitly as the cutting points of the slopes. (see Heuft et. al, 1995 for a more detailed description of this parameterization.) A pitch peak can be attributed either to a syllable or to a boundary.

The rules generate only the syllable durations. Segmental durations are calculated from the syllable durations using a formula developed by Campbell & Isard (1991).

3.2 The rule system

The rules were obtained by statistical analysis of a prosodic database, which contains about 1 hour of read speech (3 speakers) with all prosodically relevant utterance types (Heuft et. al, 1995). The speech is annotated with the model parameters described above and with linguistic and other information (e. g. the prominence values) that is supposed to influence these parameters.

The prosodic parameters are generated in the following order: syllable duration, place of F_0 peak relative to the onset of the stressed vowel, height of this peak, F_0-slope preceding and following the peak The rules are in the form of decision trees. Only binary splits are possible. Each leaf may contain either a fixed value or a linear function. The rules allow the interdependence of the parameters, e.g. syllable duration can be taken into account for the placement of the F_0-peak. The rules in the first step are mainly concerned with the position of the syllable in a prosodic phrase (after deciding if the phrase in question is progredient or terminal). In the second step, prominence of syllables and boundaries is modeled according to the findings in the previously described experiments. In the third step, adjustments are made according to the context: this step mainly applies to syllables. For example, the distance between two F_0-peaks and the prominence of adjacent syllables is considered. In the last step, microprosodic variations are taken into account, these variations are small but supposed to lead to less monotonous prosody. All input can easily be derived from the simple description mentioned above. In the next section, examples of influences on the model parameters are shown.

Before the rules can be applied, the information that is needed but not explicitly annotated in the input (position of syllables in the phrases; number and type of phones in the syllable; accentability of the syllables; type of vowel, etc.) has to be determined. Prominence values can be annotated in the orthographic input or are generated starting from the word class. The rules predict durations for all syllables and all word boundaries. F_0 peaks are assigned to syllables with a prominence >= 15 and to boundaries with a rising F_0.

4. PROSODIC PARAMETERS

4.1 Duration

Factors taken into account for syllable duration are the number of segments per syllable and the position of the syllable (phrase final or not; *final lengthening* applies to the last syllable in each phrase and also to the penultimate syllable if the last one is not accentable; i.e. contains a schwa or a vocalic /r/). The degree of final lengthening depends on the prominence of the following prosodic boundary (**Figure 1**). Another factor is the number of sonorants in the coda which increase the syllable duration. Last but not least the prominence is taken into account. There is a linear depence of syllable duration on the prominence (see Portele & Heuft, 1995).

The duration of boundaries (i.e. pauses) is 0 up to a boundary prominence of 3. For stronger boundaries, we have a linear relation beween boundary prominence and pause duration (see **Figure 2**). A value is assigned to every word boundary, so most of the boundaries have a duration of 0.

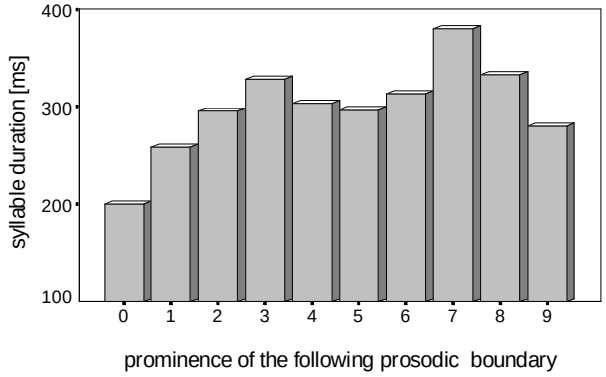

Figure 1: Duration of the phrase final syllable depending on the prominence of the following boundary. There is a linear relation up to a prominence value of 3. For stronger boundaries, prominence is coded by pause duration (see Figure 2.

4.2 Peak Delay

The correct prediction of the place of F_0 peaks (i.e. the alignment of F_0 conours with the segmental level) is perhaps the most important task for prosody generation, because of the categorical nature of pitch peak alignment that has been reported in several studies (e.g. Pierrehumbert & Steele, 1989; Kohler, 1987; Portele & Heuft, 1995). However, it is the most complex task as well. Many of factors influence on peak

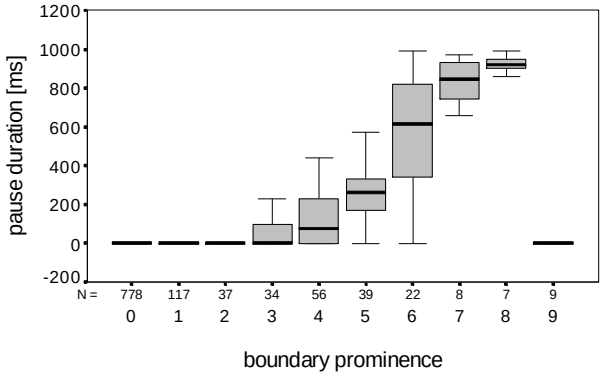

Figure 2: Duration of pauses depending on the median of the prominence ratings of three listeners. The duration at prominence 9 is 0, because 9 was usually labelled at the ends of texts.

delay.

The most important factor is the distance (in syllables) of a syllable marked with an F_0-peak from a preceding or following (for falling contours) prosodic boundary. The closer the syllable is to a following prosodic boundary, the earlier the peak is located relative to the vowel onset. In other words, speakers tend to keep the *final fall* pattern constant within certain limits (see **Figure 3**). The opposite phenomenon is found on the beginnings of phrases: The closer the syllable is to a preceding boundary, the later the peak is located relative to the vowel onset.

Not only distance to prosodic boundaries is of importance, but also the distance between two accented syllables. The closer e.g. a following syllable is, the more advanced an F_0-peak will be situated relative to the vowel onset and vice versa.

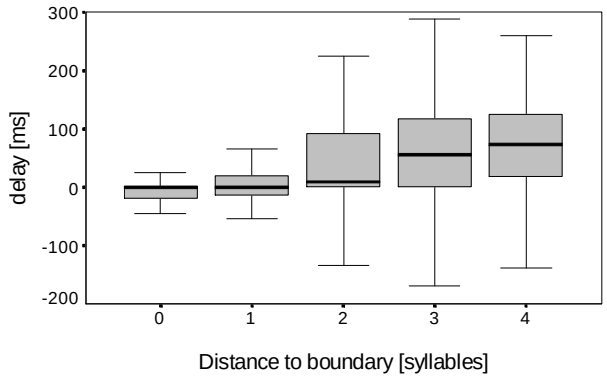

Figure 3: Distance of the F_0 peak from the onset of the accented vowel depending on the distance of the accented syllable to the following prosodic boundary.

Another factor is the segmental structure of the accented syllable. If syllable onset or nucleus contain voiced segments, the peak is often shifted in their direction (see Figure 4 as example for sonorants in the coda). This effect has been stated for Dutch by Rietveld & Gussenhoven (1995). There is also an interplay between syllable duration and location of the F_0-peak, the relation found by linear regression is: *delay= 0.2 * syllable duration*.

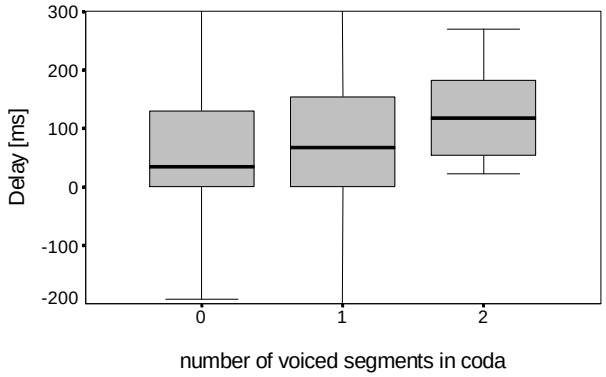

Figure 4: Distance of the F_0 peak from the onset of the accented vowel depending on the number of voiced segments in the coda. As plosives and fricatives are devoiced in syllable final position, only sonorants are considered.

4.2 Peak Height (amplitude)

Declination is not modeled by decreasing top- and baselines but by a downstep from peak to peak. This implies that the later a peak is situated within a phrase, the less high it will be. We find a relation *amplitude = -0.007 * position of the syllable in the phrase.*

Amplitude is greater when the syllable nucleus is a vowel that we found (Heuft & Portele, 1995) to have a rather high intrinsic fundamental frequency ([iː][ɪ][uː][ʊ][yː][ʏ][o][ɔ][øː][œ]).

Further, peak height depends on the prominence of the syllable (**Figure 5**): the more prominent the syllable, the higher the peak.

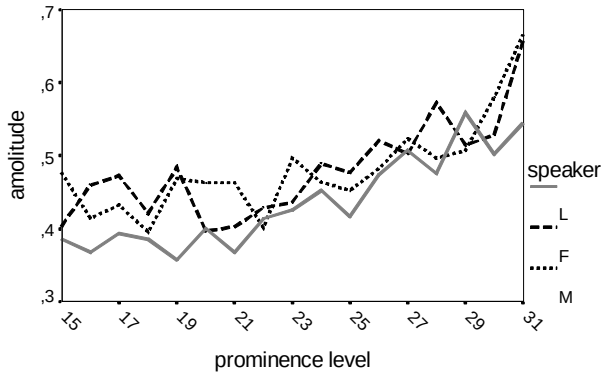

Figure 5: peak height (amplitude) against the median of prominence ratings. Only syllables with a prominence > 15 are associated with a pitch peak in our system.

Prominence of boundaries influences the height of the last F_0 peak in falling phrases. The peak is less high for strong boundaries than for weak boundaries. We could not find a relation between boundary prominence and height of the final rise for progredient utterances or questions.

4.3 Left Slope

Again, we find a strong effect of the distance of the syllable from a (preceding) prosodic boundary. If the syllable is close to a boundary, the slope is steeper (i.e. the damping factor of the cosinus curve is greater) than if the syllable is further from a boundary (see **Figure 6**).

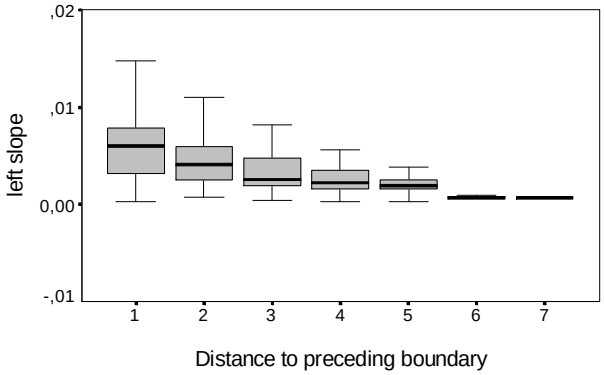

Figure 6: Steepness of the F_0 slope preceding the peak against the distance of the syllable from the preceding prosodic boundary. Higher values indicate a steeper rise.

The slope also depends on the distance (in syllables) to the preceding F_0 peaks: the smaller the distance, the steeper the fall. The height of the peak (amplitude) influences the preceding rise with a factor of 0.005.

4.4 Right Slope

Comparable to the left slope, the right slope (i.e. the steepness of the fall after an F_0 peak) depends on the amplitude of this peak (*right slope = 0.05 * amplitude*). The dependence of the right slope on the distance of the following peak is shown in **Figure 7**.

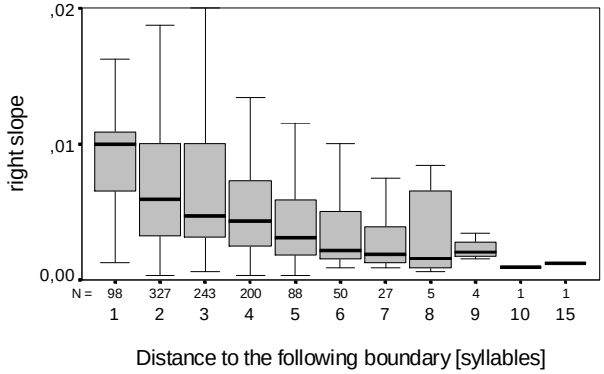

Figure 7: Steepness of the final fall (*right slope* of the last peak in a phrase) against the distance of the syllable to the following prosodic boundary.

The last right slope in phrases with falling F_0 has to be modeled very carefully because it represents the final fall. It is dependent on the distance to the following prosodic boundary, i.e. it gets steeper the nearer the syllable is situated relative to the boundary. Further, the prominence of the boundary has an effect: Stronger boundaries are modeled with steeper falls.

5. EVALUATION

5.1 Pair comparison

A simple comparison of pairs of isolated sentences generated with a) the old prosody component, and b) the tree-structured rules was carried out. The outcome was a clear preference for the new system. (All16 subjects prefering the new rules, 11 of them significantly; the overall preference was 72%). However, we must see that in such a comparison the reference system only has to be bad enough to obtain good results.

5.2 Delexicalized speech

Therefore, another experiment was carried out to compare the synthetic prosody with natural prosody (Sonntag, 1996). The pitch marks of two speakers and synthetic speech were used to produce sawtooth signals. 12 subjects were asked to recognize different syntactical and accentual structures of short declarative sentences in the sawtooth signals. The recognition rate was 80% for the natural speech and 63% for the synthetic utterances. Comparing the two natural speakers we obtained 81% vs. 67% recognition rates. Thus, we may conclude that the difference between synthetic and natural is not much bigger than inter-speaker differences.

6. DISCUSSION

In this paper we have tried to outline the rule system for prosody generation. Of course, space is too limited to present every factor influencing acoustic prosody, but the most important points have been mentioned. Many factors (e.g. prominence, boundary distance) are of a gradual nature, therefore, we often have functions rather than fixed values at the leaves of the rule trees. The functions presented here are not exactly the same as in the actual rules, because here, the interplay of the different factors is not discussed.

The evaluations gave promising results but they only considered the generation of short isolated sentences. The quality of styles that are more interesting (and more difficult to handle) like dialogues or longer texts has not been evaluated yet <a634.wav>.

It goes without saying that a good prediction of the prominence values is crucial for the resulting quality of the synthetic prosody. In a TTS system the values can be predicted from the text using information about word class, topic structure and syntax. We are aware that this is quite a difficult task. The structure that we have presented here is in fact more suitable for a concept-to-speech system.

Acknowledgement: This work has been funded by the German Federal Ministry of Education, Science, Research and Technology in the scope of the Verbmobil project under Grant 01 IV 101 G. The authors would like to thank Gerit Sonntag for the sawtooth experiment.

REFERENCES

Campbell, W.N.; Isard, S.D. (1991): Segment durations in a syllable frame. *Journal of Phonetics* **19**: 29-38

Fant, G.; Kruckenberg, A. (1989): Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* **2**, pp. 1-83

Heuft, B.; Portele, T.; Höfer, F.; Krämer, J.; Meyer, H.; Rauth, M.; Sonntag, G.: (1995): Parametric Description of F_0 contours in a Prosodic Database. *Proc. ICPhS 95, Stockholm*, pp.378-381

Heuft, B.; Portele, T. (1995): Intrinsic prosodic values and segmental context. *Proc. Eurospeech'95*, Madrid, pp. 2077-2080

Heuft, B.; Rauth, M.; Höfer, F. (1996): Prominenz von prosodischen Grenzen. To appear: *Fortschritte der Akustik - DAGA'96*.

Kohler, K. J. (1987): Categorical pitch perception. *Proc. ICPhS'87* Vol.5 Tallinn, pp. 149-152

Pierrehumbert, J.B.; Steele, S.A. (1989): Categories of tonal Alignment in English. *Phonetica* **46**, pp. 181-196

De Pijper, J.R.; Sanderman, A. (1994): On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. J. acoust. Soc. Am. **96 (4)**, pp. 2037-2047

Portele, T.; Heuft, B. (1995): Two kinds of stress perception. *Proc. of the ICPhS'95, Stockholm*, pp. 126-129

Portele, T.; Heuft, B. (1996): Towards a prominence based synthesis system. To appear: *Proc. of the 2nd Speak! Workshop*, Darmstadt

Rietveld, T.; Gussenhoven, C. (1995): Aligning pitch targets in speech synthesis: effects of syllable structure. Journal of Phonetics **23**, pp. 375-385

Sonntag, G. (1995): Klassifikation syntaktischer Strukturen aufgrund rein prosodischer Information. To appear: Fortschritte der Akustik - DAGA'96