THE MAXIMUM-BASED DESCRIPTION OF F_0 CONTOURS AND ITS APPLICATION TO ENGLISH

Thomas Portele[1] *and Barbara Heuft*[2]

[1] IKP, University of Bonn, Germany (tpo@ikp.uni-bonn.de)
[2] IKP, University of Bonn, Germany;
new affiliation: Philips Speech Processing, Aachen, Germany

ABSTRACT

The maximum-based description is a linear and simple parametrization method of F_0 contours. An F_0 maximum is characterized by four parameters: its position, its height, its left and its right slope. An automatic parametrization algorithm was developed. A perceptual evaluation was carried out for German and for English. The perceptual equality between original and parametrized contours was confirmed.

1. MOTIVATION

In 1995, we proposed a method of F0 parametrization based on an accurate description of linguistically relevant F0 maxima [7]. We aimed at a method that should

1. yield simple and intuitive parameters,
2. allow an automatic determination of the parameters,
3. result in a perceptually equal F0 contour.

We were influenced by earlier work with the model of Fujisaki [4, 12], but in our opinion that model lacked intuitivity for sake of a physiological foundation. Our primary interest was the relation between acoustic and perceptual parameters for speech synthesis [13, 15]. Besides, the parametrization should allow an intuitive formulation of rules (which is quite difficult with Fujisaki's model) and the use of a data-driven approach (which prohibited manual stylization [20]).

We therefore devised the Maximum-Based description (MBD) which is, incidentially, similar to the Tilt model [19]. The focus on maxima was influenced by Kohler [9] who showed the importance of peak position for the perception of intonation. The following three sections describe the MBD and its performance relative to the three requirements above.

2. DESCRIPTION

The MBD describes each F0 maximum by four parameters:

delay This value contains the distance from the maximum to the start of the associated vowel. It is 0 for boundary tones. Pitch accents located before the vowel have a negative delay. Delay is measured in milliseconds.

amplitude The height of a maximum is given relative to a top- and a baseline [5]. These lines are speaker-dependent and constant. Their height in Hz corresponds to the highest and lowest frequencies of the modal voice of a given speaker. If the peak reaches the top line, the amplitude value is 100. Although amplitude is dimensionless, its value is computed using the Hz scale.

rise The rise towards a maximum is approximated by $\cos^2$ curves. A curve begins on the baseline and ends at the maximum. The distance V in ms between curve start and curve end is used to compute the rise value: $rise = \frac{N \cdot amplitude}{V}$, N is a scaling factor which is set to 4 in order to obtain average values between 0 and 1.

fall The fall value is computed like the rise value.

Minima are not explicitely modelled but are located at the place where a falling line meets the next rising line. Thus, connection elements like those used in the Tilt model [19] are not necessary. Figure 1 gives an example.

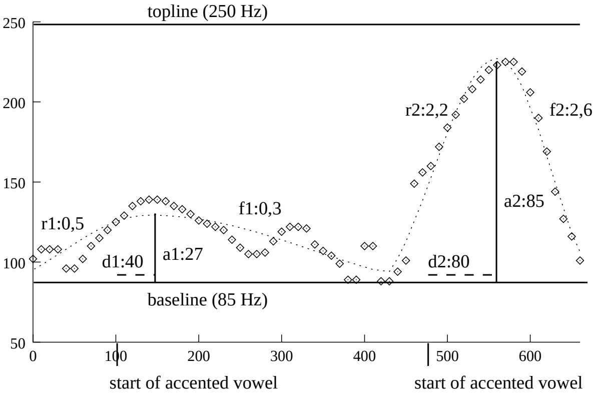

Figure 1. Original (squares) and parametrized contour (dotted line) of the utterance "Where have you been?" with accents on "Where" and "been". Displayed are the delays $d1$ and $d2$, the amplitudes $a1$ and $a2$, the rises $r1$ and $r2$ and the falls $f1$ and $f2$. The minimum is at the intersection of two $\cos^2$ curves.

The MBD allows a simple description with intuitive parameters. It is possible to synthesize acceptable F0 contours by a few simple rules written from scratch [14].

In an earlier version top- and baseline were not kept constant but were adapted for each utterance (the MBD thus was a superpositional model). Here, interactions between accents and lines appeared; an initial maximum can be described by a high baseline start or by a maximum with a large amplitude. Furthermore, additional degrees of freedom were present. After informal experiments that showed a perceptually equal parametrization for fixed and for variable lines, we decided to keep both lines constant and fixed.

The temporal alignment relative to the start of the associated vowel was chosen because, unlike P-center positions [11], this value is simple to obtain and often used [10, 19]. The height is calculated in the linear Hz domain. This is also done for practical reasons; the linear scale is no central point of the MBD. The same holds for the description of slopes by $\cos^2$ curves. The decisions over an operationalization of the MBD were based on pragmatic reasons.

3. PARAMETRIZATION

An automatic parametrization algorithm was constructed [16]. The number and approximate location of the maxima (i.e. their association with the segmental string) must be assigned beforehand. This can be done automatically [18] but in our corpora [7, 3] we preferred hand annotation in order to guarantee that only linguistically relevant maxima are described.

The algorithm computes the parameters for each maximum from left to right. Input values are n F0 values $F0_1, F0_2, \ldots, F0_n$ at time points $t_1, t_2, \ldots, t_n$ and the positions of the starts of the vowels asscociated with m annotated maxima $vpos_1, vpos_2, \ldots, vpos_m$. The algorithm uses five steps

1. The time of the k-th maximum $maxindex_k$ in the vicinity of the k-th accented vowel $vpos_k$ is established.
2. The delay is computed:

$$delay_k = t_{maxindex_k} - vpos_k$$

3. The amplitude is computed relative to topline and baseline:

$$amplitude_k = 100 \frac{F0_{maxindex_k} - baseline}{topline - baseline}$$

4. The optimal rise parameter is calculated with a simple optimization procedure for those F0 values between the last minimum and the current maximum. The optimization criterion is the minimization of the sum of square errors; a linear weighting function ω_i increases error values near the maximum in order to obtain a better fit in this region:

$$\omega_i = \left(1 + \omega \frac{t_i - t_{minindex_{k-1}}}{t_{maxindex_k} - t_{minindex_{k-1}}}\right)$$

$$height_i = (topline - baseline) \cdot \frac{amplitude_k}{100}$$

$$cosval_i = \cos^2\left(\frac{(t_i - t_{maxindex_k}) \cdot rise}{4 \cdot amplitude_k}\right)$$

$$F0fit_i = (baseline \cdot height_i \cdot cosval_i)$$

$$error(rise) = \sum_{i=minindex_{k-1}}^{maxindex_k} \omega_i \cdot |F0_i - F0fit_i|$$

$$rise_k = min(error(rise))$$

Here, $minindex_{k-1}$ is the value of the $(k-1)$-th minimum (or the first value if $k = 1$), and ω ist the weighting parameter; if $\omega = 0$, no weighting is performed.

5. The optimal fall parameter is calculated by an analogous procedure. If no further maximum exists, $minindex_k$ is the last F0 value, else it is the lowest F0 value between the current and the next maximum:

$$\omega_i = \left(1 + \omega\left(1 - \frac{t_i - t_{maxindex_k}}{t_{minindexk} - t_{maxindex_k}}\right)\right)$$

$$height_i = (topline - baseline) \cdot \frac{amplitude_k}{100}$$

$$cosval_i = \cos^2\left(\frac{(t_i - t_{maxindex_k}) \cdot fall}{4amplitude_k}\right)$$

$$F0fit_i = (baseline + height_i \cdot cosval_i)$$

$$error(fall) = \sum_{i=maxindex_k}^{minindex_k} \omega_i \cdot |F0_i - F0fit_i|$$

$$fall_k = min(error(fall))$$

If unsmoothed F0 contours are parametrized, minima and maxima have to be assigned manually. Figure 2 displays an example of a parametrized contour.

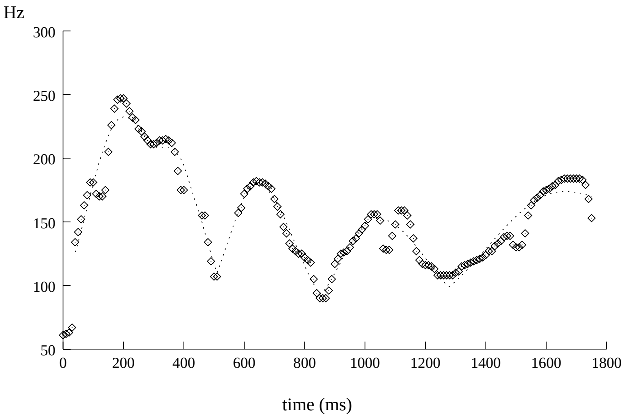

Figure 2. Automatic parametrization (dotted line) of an F0 contour (squares) of the utterance “Would you like to come with me to the movies”.

4. EVALUATION

The goal of the MDB is the computation of a parametrized contour that is “perceptually equal” [20] to the original contour. Two different experiments were carried out, one for German and one for English. Both experiments are perception tests, because no numerical value can reliably predict perceptual equality due to

- the limited amout of knowledge about pitch perception of natural speech,

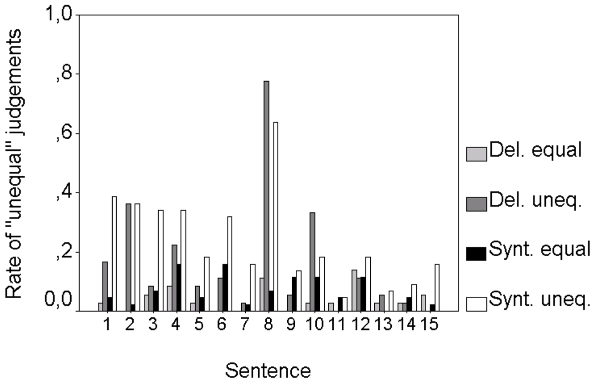

Figure 3. Result of the perception tests for German. The rate of "unequal" judgements is displayed for resynthesized and delexicalized equal and unequal pairs for each sentence. Sentence 8 had a flat contour.

- the goal of the stylization; a stylized contour should omit as many irrelevant details as possible. A successful stylization may have a large numerical difference (caused by the omission of irrelevant details) while remaining perceptually equal.

The explicit neglection of the F0 contour at minima by the MBD causes sometimes rather large numerical discrepancies without any audible difference.

4.1. German

Fifteen utterances varying in length and prosodic structure were recorded, automatically parametrized, and resynthesized [8]. One of these resynthesized versions had a flat contour without any maxima in order to assess whether subjects discriminate at all. Eleven subjects judged pairs OO, OP, PO, and PP (O = Original, P = resynthesized version) if they were the same or not. The pairs OP, PO were judged more often as "different" (21 %) than the OO, PP versions (7 %). The subjects commented that audible distortions induced by the resynthesis process made it difficult to concentrate solely on prosody. Therefore, the stimuli were delexicalized using the PURR method [17]; the test was repeated with delexicalized stimuli and nine subjects (Figure 3).

The differences between OP,PO and OO,PP judgements were significant only for three of the fifteen sentences. In one case of a very short stimulus ("Nein" *No*) an error in the delexicalization process caused an audible crack (sentence 10). The second case was the flat contour (Sentence 8), and in the third case a parametrization error was caused by a bug in the algorithm (Sentence 2). In all other cases, original and parametrized versions were perceptually indistinguishable.

4.2. English

Five utterances were randomly chosen from a large corpus [3]. These were parametrized, resynthesized, and delexicalized. Four versions of each utterance were constructed:

1. the original utterance (O),
2. the automatically parametrized utterance (P),
3. an utterance where maxima were placed correctly but each maximum shared the same default parameters (D),
4. a flat contour with no maxima (F).

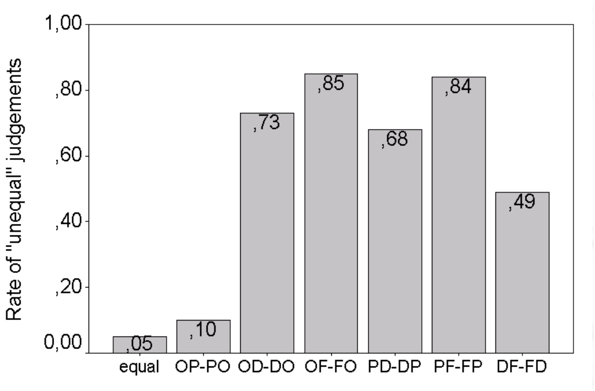

Figure 4. Result of the perception tests for English. The rate of "unequal" judgements is displayed for all pairs (see text).

Each utterance was presented to eight subjects in pairs OO, PP, DD,FF, OP, PO, OD, DO, OF, FO, PD, DP, PF, FP, DF, FD. The subjects' task was the same as in the previous experiment.

All subjects were native German speakers. One might argue about the value of this experiment because no native listeners participated. We assume that, unlike tests for "perceptual equivalence" or "functional equality" or similar rather diffuse categories where linguistic knowledge may become important, "perceptual equality" does not depend on the native language if languages as similar as English and German [6] are concerned. In the German experiment, one subject was a native speaker of English, and his performance was similar to those of the German listeners. The German and English stylizations of the IPO researchers, who are native Dutch speakers [2, 1], are judged as perceptually equal by native listeners. Investigator and subjects had an identical notion of "perceptual equality". We therefore assume that perceptual equality of American utterances (especially delexicalized ones) can be assessed by German listeners.

The results (Fig. 4) show that the subjects could distinguish all pairs except those with identical versions and the OP-PO pairs. The automatic parametrization yields contours that are perceptually equal to the original contours. The shape of a maximum is perceptually relevant: contours OD-DO and PD-DP are perceived as "different". The results indicate that

- the delexicalization prevails the intonation contour,
- the original and parametrized version are perceptually equal,
- the choice of MBD parameters is perceptually salient.

4.3. Numerical Analysis

A numerical analysis was performed for the German stimuli. As discussed above, mean distance between original and parametrized stimuli is no useful measure. The correlation coefficient is better suited to capture overall similarities. The mean correlation coefficient for all 15 sentences is 0.85. The correlation coefficient between the individual coefficients for each sentence and the corresponding number of "unequal" judgements is -0.89 for the delexicalized stimuli; high correlations are likely to predict a low number of "unequal" judgements. The same value for the resynthesized stimuli is -0.59 which indicates that delexicalization aids the perception of intonational similarities.

5. DISCUSSION

The MBD allows the automatic parametrization of F0 contours (unlike the manual IPO stylization [20]). The parametrization was found to be perceptually equal to the original contour. Only a few intuitively meaningful parameters are sufficient (unlike the physiologically motivated parameters of the Fujisaki model that are more difficult to interpret), and only one unit, the maximum, is necessary (unlike the Tilt model [19] which makes another kind of simplification in combining left and right slope in one parameter). It is unlikely that a further reduction of parameters is possible. Amplitude is essential for the perception of declination and the marking of extra prominence [15]. Delay is necessary to align the F0 contour with the segmental string. Although this alignment is not evaluated when using delexicalized stimuli, the rhythmic pattern prevails and distortions are perceivable. If minima are judged as relatively unimportant by the MBD they have to be modeled, and this can only be done by adjusting left and right slope. A parametrization based on a linear sequence of pitch accents and boundary tones must contain at least those four parameters. The relevance of the four parameters is further demonstrated by the performance of the parametrized versions with default parameters in the evaluation.

By successfully generating perceptually equal parametrizations we have shown that minima or low targets are not as important as maxima regarding their phonetic manifestation, because the difference between original and parametrized versions is larger in regions with low F0 without destroying perceptual equality. This may be due to the fact that these regions usually also have low energy. In our interpretation, minima have no clear linguistic function, save the final fall at the end of a statement. This should be modeled by special parameters, perhaps not necessarily in the parametrization process, but the generation of satisfying synthetic F0 contours is much easier [14].

6. ACKNOWLEDGEMENT

We thank Jürgen Krämer for his assistance in programming the automatic parametrization algorithm, Barbertje Streefkerk for her help with the German experiment, Gerit Sonntag for PURRing the stimuli, and all subjects for their patience.

7. REFERENCES

[1] L. M. H. Adriaens. *Ein Modell deutscher Intonation*. PhD thesis, Technische Universiteit Eindhoven, 1991.

[2] J. R. de Pijper. *Modelling British English Intonation*. Foris, Dordrecht, 1983.

[3] A. Elsner, M. Wolters, T. Portele, M. Rauth, and G. Sonntag. Designing and labelling a prosodic database for American English. In *Proceedings of the Workshop on Language Resources, Granada*, 1998.

[4] H. Fujisaki. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In O. Fujimura, editor, *Vocal physiology: voice production, mechanisms and functions*, pages 347–355. Raven, New York, 1988.

[5] E. Gårding. A generative model of intonation. In A. Cutler and D. R. Ladd, editors, *Prosody: Models and Measurements*. Springer, Berlin, 1983.

[6] E. Grabe. Pitch accent realization in English and German. *Journal of Phonetics*, 26:129–143, 1998.

[7] B. Heuft, T. Portele, F. Höfer, J. Krämer, H. Meyer, M. Rauth, and G. Sonntag. Parametric description of F0-contours in a prosodic database. In *Proceedings of the XIII. International Congress of Phonetic Sciences, Stockholm*, volume 2, pages 378–381, 1995.

[8] B. Heuft, B. Streefkerk, and T. Portele. Evaluierung der automatischen Parametrisierung von Grundfrequenzkonturen. In *Tagungsband Elektronische Sprachverarbeitung VII, Berlin*, pages 170–175, 1996.

[9] K. J. Kohler. Categorical pitch perception. In *Proceedings of the XI. International Congress of Phonetic Sciences, Tallin*, volume 4, pages 331–333, 1987.

[10] K. J. Kohler. Prosody in speech synthesis: the interplay between TTS and basic research. *Journal of Phonetics*, 19:121–138, 1991.

[11] S. M. Marcus. Acoustic determinants of perceptual centre (p-centre) location. *Perception and Psychophysics*, 30:247–256, 1981.

[12] B. Möbius. *Ein quantitatives Modell der deutschen Intonation — Analyse und Synthese von Grundfrequenzkonturen*. Niemeyer, Tübingen, 1993.

[13] T. Portele. Perceptual evidence for accent categories: preliminaries and first results. In *Proceedings of the ESCA Workshop on Intonation, Athens*, pages 271–274, 1997.

[14] T. Portele. Eine perzeptiv motivierte Beschreibung der Intonation des Deutschen. In *Tagungsband Elektronische Sprachverarbeitung IX, Dresden*, (to appear), 1998.

[15] T. Portele. Perceived prominence and acoustic parameters in American English. In *Proceedings of the International Conference on Spoken Language Processing, Sydney*, 1998.

[16] T. Portele, J. Krämer, and B. Heuft. Parametrisierung von Grundfrequenzkonturen. In *Fortschritte der Akustik - DAGA 95, Saarbrücken*, pages 991–994, 1995.

[17] G. P. Sonntag and T. Portele. PURR – a method for prosody evaluation and investigation. *Computer Speech and Language*, 12 (to appear), 1998.

[18] V. Strom. Detection of accents, phrase boundaries and sentence modality in German with prosodic features. In *Proceedings of the Eurospeech 95, Madrid*, pages 2039–2041, 1995.

[19] P. Taylor and A. W. Black. Synthesizing conversational intonation from a linguistically rich input. In *Proceedings of the IEEE/ESCA Workshop on Speech Synthesis, New Paltz*, pages 175–178, 1994.

[20] J. t'Hart, R. Collier, and A. Cohen. *A perceptual study of intonation*. Cambridge University Press, Cambridge, 1990.