Data-driven Extraction of Intonation Contour Classes

Uwe D. Reichel

Institute of Phonetics and Speech Processing
University of Munich, Germany
reichelu@phonetik.uni-muenchen.de

Abstract

In this paper we introduce the first steps towards a new data-driven method for extraction of intonation events that does not require any prerequisite prosodic labelling. Provided with data segmented on the syllable constituent level it derives local and global contour classes by stylisation and subsequent clustering of the stylisation parameter vectors. Local contour classes correspond to pitch movements connected to one or several syllables and determine the local f0 shape. Global classes are connected to intonation phrases and determine the f0 register. Local classes initially are derived for syllabic segments, which are then concatenated incrementally by means of statistical language modelling of co-occurrence patterns.

Due to its generality the method is in principal language independent and potentially capable to deal also with other aspects of prosody than intonation.

1. Introduction

The prosody module of a speech synthesis system has to relate text or concepts to prosody in order to predict the latter from the former. To facilitate this mapping some representation of prosody is needed. Since this paper deals with the intonational aspect of prosody, some common description approaches for intonation are shortly listed here. They can roughly be divided into symbolic, parametric and perception-based approaches.

1.1. Symbolic Approaches

In the Tone Sequence Approach [19], which is grounded on auto-segmental phonology [22], intonation is seen as a succession of tones that are associated to accentuated or phrase final syllables. The tone inventory consists of two elementary tones (High and Low) that can be combined to complex tones. Possible tone sequences are controlled by an intonation grammar. There are rule-based [20] statistic approaches [21] for the generation of the concrete f0 values from this abstract representation of intonation.

The Kiel Intonation Model (KIM) [6] treats prosodic categories as bundles of distinctive features. It contains rules for mapping manual annotations to prosodic categories, and for mapping those categories to numeric f0 values. One emphasis lies on examining the synchronisation of syllable nuclei and f0 peaks (so called *early, middle and late peak*).

1.2. Parametric Approaches

The Fujisaki model ([8], [10], [11]) predicts intonation contours by a superposition of a baseline f0, a phrase component for global contours (intonation phrases), and an accent component for local contours (accented syllables). One possibility to estimate this model's parameter values is analysis by synthesis [11], i.e. analysing the given f0 contour by synthesis via the Fujisaki model.

Models like Tilt [12] and PaintE [7] try to approximate the f0 contour on accentuated syllables by stylisation functions. In PaintE furthermore the parameter vectors of the stylisation function are clustered in order to get categorised intonation building blocks.

1.3. Perception-based Approaches

The IPO model ([24], [23]) operates on a perceptually equivalent approximation of given f0 contours by a sequence of straight lines (the so called *copy contour*). Thus this stylisation is carried out interactively with subjects judging the approximation perceptually. The resulting lines of different slope form intonation units who's succession can be described by an intonation grammar.

1.4. Shortcomings of the Given Approaches

There are some shortcomings of the approaches described above:

- Leaving aside IPO, all models mentioned above rely on accent and phrase boundary labels of various complexity. Therefore at least initially hand-labelling of the data is necessary. This work is time consuming and needs trained experts. Especially in prosody inter-labeller agreement and intra-labeller consistency run the risk of getting relatively low [4] which leads to a loss of prosodic training data. Presumably this problem grows with the increasing size of the label inventory.

- The label inventories are not necessarily language independent. Inventories like ToBI for example need to get adjusted whenever they are applied to new languages [5]. Also the IPO model needs perceptual readjustment for each new language.

With our model we try to avoid these shortcomings. Since our approach is purely data driven, no manual prosodic labelling or manual adjustment to other languages is needed.

2. Data

Our training data consists of parts of the IMS Radio News Corpus [1] with a total length of about 14 minutes. The corpus part used in this study contains news texts read by one professional male speaker. It is segmented amongst others on the phone and syllable level. For f0 measurement we utilised autocorrelation implemented in *Praat* (version 4.1.5) software with a sampling rate of 100 Hz.

3. Extraction of Local Contour Classes

As in the Fujisaki model, our model distinguishes between local and global contours. Local contour classes correspond to pitch shapes connected to one or more syllables. They are derived by parameter clustering of stylisation polynomials. Starting with syllables, contour segments are iteratively merged to larger units. Figure 1 gives an overview over the processing steps which are described in greater detail in the following sections.

```
segments := syllables
iterate
    foreach $s \in$ segments
        - preprocessing: interpolation, smoothing, and time
          normalisation of f0 contour of s in context of the pre-
          ceeding and following syllable.
        - adaptive stylisation of the contour by polynomials
    end
    - cluster polynomial coefficients to derive contour classes
    - segments := merge neighbouring segments if respective
      classes occur in dependence of each other
    - terminate if no merging possible
end
```

Figure 1: *Algorithm for incremental local intonation contour class extraction*

3.1. Contour Preprocessing

Preprocessing as shown in Figures 2 and 3 removes contour characteristics not related to intonation, among them microprosody, intrinsic pitch, speech rate, and syllable constituency. For each contour segment preprocessing took place in the context of the preceeding and the following syllable. Hertz values were transformed to the logarithmic semitone scale.

3.1.1. Smoothing

To eliminate f0 movements related to intrinsic pitch, coarticulation effects at voice on- and offsets and f0 measurement errors the contours were smoothed using a Savitzky-Golay filter [18] of order 3 and length 5. This filter is commonly applied for this purpose (see e.g. [17]) due to its capability to remove high frequency noise from pertinent information.

3.1.2. Time Normalisation

In order to exclude any influence of speech rate, phone number, and syllable constituent structure, all syllables were time normalised in the following way: the syllable head is mapped on the interval -0.4 to -0.2, the nucleus from -0.2 to 0.2, and the coda from 0.2 to 0.4. Missing heads or codas are padded by interpolation between the f0 values of the nucleus and the neighbouring syllables.

3.1.3. Interpolation

Since the subsequent stylisation step requires continuous contours, plosive closure phases and missing syllable constituents are bridged by cubic splines.

3.2. Adaptive Stylisation

A polynomial stylisation was carried out, guided by the multidimensional unconstrained nonlinear Nelder-Mead minimisation [15] of the squared error between original and stylised contour.

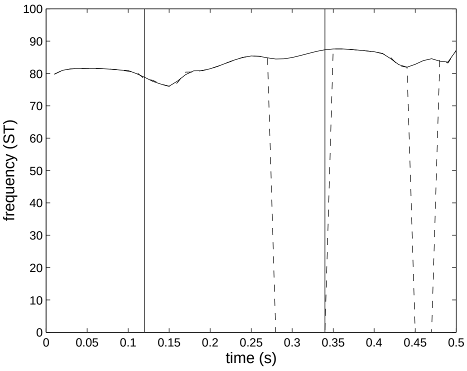

Figure 2: *Preprocessed f0 contour (solid line): spline interpolation, smoothing by Savitzky-Golay filter. Dashed line: original contour. The two vertical lines mark the boundaries of the contour segment and the preceeding and following syllable, respectively.*

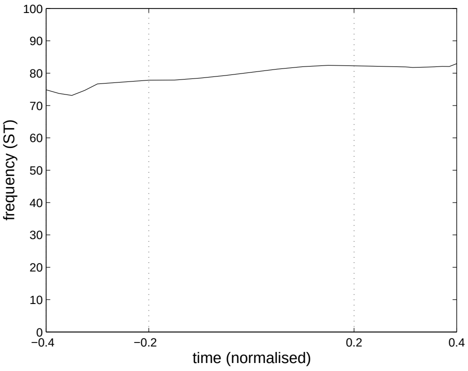

Figure 3: *Time normalisation of the f0 contour. The vertical lines separate syllable onset, nucleus and coda.*

The higher the polynomial order, the closer the fit to the original contour, but also the more unreliable the subsequent clustering of the coefficient vectors. Therefore for each contour stylisation the lowest possible polynomial order was chosen ranging from zeroth to third order (cf. Figure 4). The goodness of fit was determined by the maximum distance between corresponding values of the original and the stylised contour. If this distance did not exceed a certain value, the stylisation was judged to be sufficiently close. The threshold was set to 4 Hz with reference to Klatt [16] who reported a just noticeable difference of 2 to 5 Hz for non-stationary stimuli.

To make sure that all coefficient vectors had the same length for subsequent clustering, zeros were padded to the vectors of the polynomials of lower order than 3.

3.3. Clustering

As in the PaintE model mentioned in the introduction, intonation contour classes were derived by Kmeans clustering of the coefficient vectors of the stylisation polynomials. Since only the shape and not the frequency offset characterises a contour class, the first coefficient was ignored.

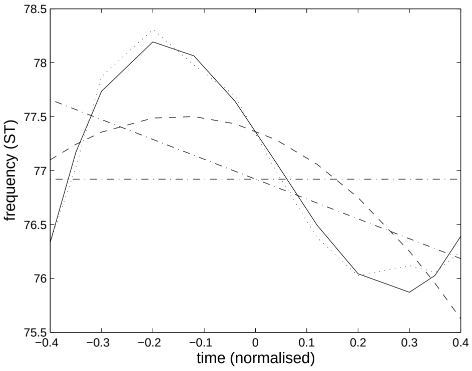

Figure 4: *Adaptive stylisation using polynomials of increasing order until maximum distance criterion is met. Dotted line: contour to be stylised, remaining lines: polynomial stylisation of increasing order from 0 to 3.*

Here the determination of the optimal number of clusters was guided by the Dunn index, a validity measure of hard clustering taking into account cluster compactness and separation between clusters. After having carried out Kmeans clustering 10 times for each given specification of number of clusters, the number connected to the highest mean Dunn score was chosen.

The centroid vectors served as cluster representatives.

4. Merging of Contour Segments

Contour segments are merged if the respective contour classes co-occur non-randomly. To determine whether the co-occurrence is random or not, the Log-Likelihood Ratio is utilised, a method used in the field statistical natural language processing for example to retrieve collocations [13].

This method compares the likelihoods L of the observed occurrences of the intonation classes c_i and c_j given two different hypotheses:

$$\begin{array}{ll} \text{H0:} & P(c_i|c_j) = p = P(c_i|\neg c_j) \\ \text{H1:} & P(c_i|c_j) = p_1 \neq p_2 = P(c_i|\neg c_j) \end{array}$$

According to $H0$, c_i and c_j occur independently (the probability p of c_i does not change in dependence of preceeding c_j), whereas $H1$ claims dependence. Under the simplifying assumption that the probabilities for the observed occurrence pattern for c_i and c_j can be described by a binomial distribution, the likelihoods for the observed data according to $H0$ and $H1$ are given as follows:

$$\begin{aligned} L(H0) &= b(n_{ij}; n_j, p) b(n_i - n_{ij}; N - n_j, p) \\ L(H1) &= b(n_{ij}; n_j, p_1) b(n_i - n_{ij}; N - n_j, p_2), \end{aligned}$$

where n_i and n_j are the observed frequencies of classes c_i and c_j, respectively, n_{ij} stands for the observed frequency of the sequence $c_i c_j$, and N is the total number of observations. The probability $b(n_{ij}; n_j, p)$ following a binomial distribution then represents the expectation of observing the sequence $c_i c_j$ n_{ij} times in n_j trials, if the probability of observing c_i given c_j is p.

A comparison of the log likelihoods leads to the Log-Likelihood Ratio $\ln \lambda$:

$$\begin{aligned} \ln \lambda &= \ln \frac{L(H0)}{L(H1)} \\ &= \ln L(n_{ij}, n_j, p) + \ln L(n_i - n_{ij}, N - n_j, p) \\ &\quad - \ln L(n_{ij}, n_j, p_1) - \ln L(n_i - n_{ij}, N - n_j, p_2) \end{aligned}$$

$-2 \ln \lambda$ follows approximately a χ^2 distribution, so a χ^2 test can be applied to decide whether the independence hypothesis $H0$ can be rejected in favour of $H1$. If the dependence hypothesis turned out to be significantly more appropriate (this study's significance level was set to 0.01), the corresponding segments were merged.

The next iteration step's preprocessing, stylisation and clustering then operated on the resegmented data. In case of impossibility of further merging the procedure terminated.

5. Extraction of Global Contour Classes

As explained above, local intonation contour classes were derived independently of registers. In order to model the f0 register of each syllable, global contour classes were extracted by stylisation and clustering of f0 baselines in intonation phrases which had been segmented automatically.

5.1. Segmentation of Intonation Phrases

The baseline f0 values served as a representation of registers. For each syllable such a baseline value was calculated by taking the mean of the n lowest f0 values measured within the syllable (n was set to 8 in this study). The mean was taken to reduce the effect of potential pitch measurement errors.

As shown in Figure 5 we then simply treated each speech pause and each baseline pitch discontinuity as a phrase boundary. The discontinuity threshold was set to 3 semitones. This is only a first approximation, since also prominent pitch accents and boundary tones show that large pitch differences compared to the neighbouring syllables.

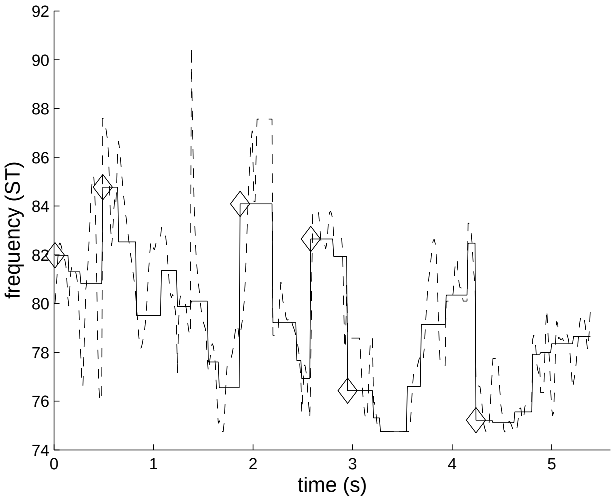

Figure 5: *Dividing the utterance into intonational phrases at baseline pitch discontinuities and speech pauses. Dashed line: original contour, solid line: pitch baseline. Intonation phrase starting points are marked by diamonds.*

5.2. Stylisation and Clustering

The time of each intonation phrase was normalised to the interval $[0\ 1]$ in order to remove any phrase length effects. The sequence of baseline f0 values of the contained syllables was stylised by straight lines, and the slope parameters were clustered by the same procedure as described in section 3.3. This led to a limited number of discrete global contour classes. As with local contour classes the centroids were taken as cluster representatives.

6. Resynthesis

In resynthesis the original f0 values were replaced by contours derived from the respective local and global contour classes. The local class determined the shape, the global class, together with the position of the segment in the intonation phrase, determined the register. An illustrative example is given in Figure 6. There the f0 contour of the one-syllable segment belongs to the local intonation class l_3 whos representative is the centroid parameter vector $p_{l3} = [16.0663, -16.1095, -88.2260]$ and to the global intonation class g_3 represented by the centroid shape parameter $p_{g3} = 6.7543$ (cf. Figures 7 and 8, respectively). The local contour $\mathrm{f0}_l$ of the time normalised segment (see section 3.1.2) is given by:

$$\mathrm{f0}_l(t) \quad = \quad 16.0663t - 16.1095t^2 - 88.2260t^3,$$

t stands for (normalised) time. The syllable dependent register $\mathrm{f0}_r(\sigma_n)$ is derived from the slope of the global contour line associated with class g_3 and the relative position $p(\sigma_n)$ of the n-th syllable σ_n within the intonation phrase. The starting point of the straight line $\mathrm{f0}_r(\sigma_1)$ is set to the original intonation phrase‘s initial baseline value. Future research is needed to predict this value, which reflects the amount of pitch reset at phrase boundaries. The registers for all syllables σ_n are then calculated the following way:

$$\mathrm{f0}_r(\sigma_n) \quad = \quad \mathrm{f0}_r(\sigma_1) + 6.7543p(\sigma_n)$$

In our example $\mathrm{f0}_r(\sigma_1)$ is 80 semitones (ST), and $p(\sigma_n)$ is 0.8 (e.g. $n = 8$ in a 10-syllable phrase). For each syllable σ_n involved in the contour segment the baseline $b_l(\sigma_n)$ of the corresponding part of the local contour $\mathrm{f0}_l$ is then replaced by the syllable's register $\mathrm{f0}_r(\sigma_n)$ so that the actual contour f0 is calculated by:

$$\mathrm{f0}(t) \quad = \quad \mathrm{f0}_l(t) - b_l(\sigma_n) + \mathrm{f0}_r(\sigma_n)$$

Finally the resulting contour is aligned to the given time range and syllable structure of the segment.

To enhance naturalness of the resulting signals we added jitter in form of a quasi-random component Δf_0 as a sum of three sine waves according to a formula proposed by Klatt and Klatt [25]:

$$\Delta \mathrm{f0}(t) \quad = \quad \frac{fl}{50} \cdot \frac{\mathrm{f0}}{100} \big[\sin(2\pi 12.7t) + sin(2\pi 7.1t) + sin(2\pi 4.7t) \big] \,\mathrm{Hz}$$

The fluttering parameter fl was set to 25. Time t is given in seconds.

7. Perceptual Evaluation

In order to test the perceptual appropriateness of our model we conducted two perception experiments, one for naturalness judgements, and the second to test functional equivalence of original and modelled contours. The stimuli were created by MBROLA (version 3.01h) resynthesis [14] replacing the original f0 contour by a sequence of contour classes as described in section 6.

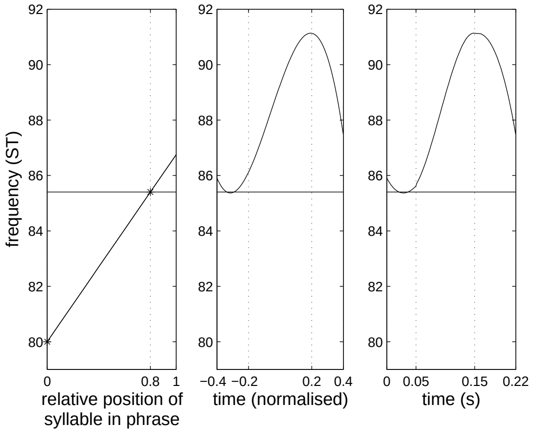

Figure 6: *Combination of global and local contour.* **Left:** *The segment's register baseline is predicted by original frequency offset (here: 80 ST), global contour associated to corresponding global contour class (here: class g_3, cf. Figure 8), and relative position of the segment within the intonational phrase (here: 0.8).* **Middle:** *The baseline value (cf. section 5.1) of the local contour given here by local contour class l_3 (cf. Figure 7) is shifted to this value.* **Right:** *The contour is aligned to the original time ranges of syllable onset (0s–0.05s), nucleus (0.05s–0.15s) and coda (0.15s–0.22s).*

6 subjects, 2 male and 4 female, took part in the experiments, their age ranged from 24 to 50. All except one were trained phoneticians, and all except one were German native speakers (the non-native speaker has lived in Germany for more than 15 years, and her pronunciation showed no foreign language accent).

7.1. Naturalness

In the first experiment the subjects were instructed to judge the naturalness of 50 inter-pausal speech segments that comprised at least 3 syllables. Each segment was presented with original and modelled f0 resulting in 100 stimuli that were randomly ordered. The judgement scale contained 4 values: *completely natural, tolerably natural, rather unnatural*, and *completely unnatural*. Since all stimuli were created using MBROLA, none of the stimulus groups was penalised compared to the other concerning synthesis artefacts. The stimuli were faded in and out by superimposing a Tukey window (taper sections each set to 3% of the stimulus length).

The participants could listen to each stimulus as often as they wanted to and could revise their judgements at any time.

Table 1 shows the mean judgements for original and modelled f0.

7.2. Functional Equivalence

In the second experiment the subjects had to decide for stimulus pairs whether their intonation contours were functionally equivalent or not. The same 50 stimuli as in the first experiment were used and presented pairwise in random order. In half of the stimuli pairs both stimuli contained either the original or the modelled f0 (‘same contour’ case). In the other half one stimulus contained the original f0, and the other the modelled one (‘different contour’ case), original and model presented in random order.

Functional equivalence concerned weighting of information (position and prominence of accents), discourse embedding of the segment (progredient vs. final intonation contour) and, if applicable, sentence mode.

As in the first experiment, each stimulus could be listened to arbitrarily often, and judgements could be revised at any time.

8. Results

8.1. Resulting Contour Classes

The application of our method to the given data yielded 9 local intonation contour classes (see Figure 7) differing in shape and number of involved syllables (from 1 to 3), and 6 global contour classes differing in slope of the declination and inclination baselines (see Figure 8).

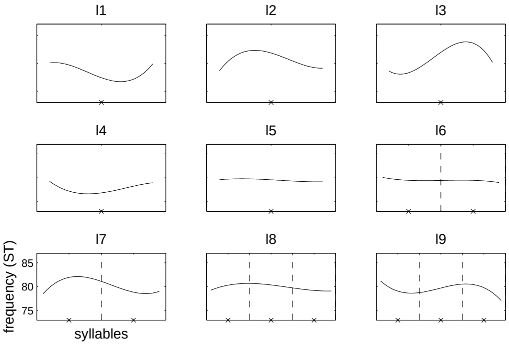

Figure 7: *Local contour classes. All contours are shifted to the mean of 80 ST. Time is normalised as described in section 3.1.2. Syllable boundaries are marked by vertical dashed lines, nucleus centers by crosses.*

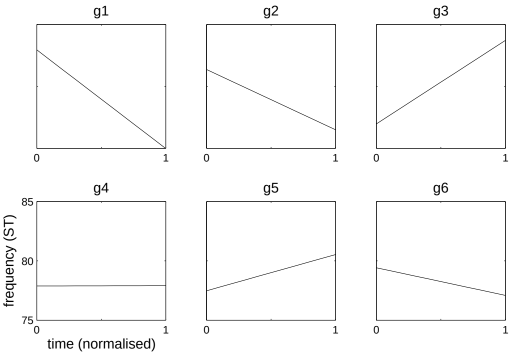

Figure 8: *Global contour classes (declination baselines). Time is normalised to the interval* [0 1].

8.2. Numerical and Perceptual Evaluation

The root mean square error between all original and generated f0 values amounted 10.26 Hz.

The results of naturalness and functional equivalence judgements are shown in Tables 1 and 2, respectively. The original f0 contours were judged highly significantly as more natural then the modelled contours (two-sided Wilcoxon matched pairs signed rank test, $alpha = 0.001$).

Table 1: *Mean subject judgements for the naturalness of original and modelled f0 contours.*

	mean	maximum	minimum
original	3.14	4	1
model	2.61	4	1

Concerning functional equivalence, Table 2 reveals that about 27% of the ‘different contour’ stimulus pairs were also judged as different. Figure 9 gives the numbers of ‘functionally not equivalent’ judgements for each of the ‘different contour’ stimulus pair types.[1]

Table 2: *Contingency table for ’functionally equivalent/not equivalent‘ judgements of stimulus pairs with same and different contours. Cramer‘s V = 0.38.*

	same contours	different contours
equivalent	149	109
not equivalent	1	41

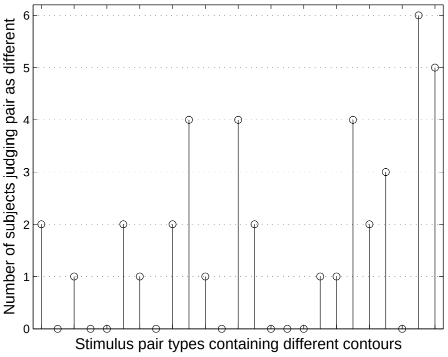

Figure 9: *Number of subjects with ‘functionally not equivalent’ judgements for each ‘different contour’ stimulus pair type.*

9. Discussion

9.1. Evaluation Results

The results of the naturalness experiment presented in the previous section clearly show that our model in its current state is not capable to produce contours that reach the quality of original intonation. This is not surprising since purely data driven models lack expert knowledge included into the models listed in the introduction, for example knowledge about perceptual equivalence (IPO) or position and types of accents and phrase boundaries. It is unclear whether the mean naturalness judgements for our model would rise, if the subjects would compare them not only to the original contours but also to a model worse than ours. Such a triple comparison had been carried out e.g. by Möhler [2] who additionally had presented flat intonation contours. Furthermore, as with all data driven models more training data is likely to enhance performance, so far we use just 14 minutes of speech.

Concerning functional equivalence, the results are already a bit more promising. As can be seen in Figure 9, 24% (6 out

[1] By stimulus pair *types* we mean the set of distinct stimulus pairs.

of 25) of the ‘different contour’ stimulus pair types were judged as functionally not equivalent by half of the subjects or more, indicating that the majority of the subjects was not able to functionally distinguish the other 76%.

9.2. Local Contours

Some of the extracted local contour classes can be related to other intonation description systems. Thus, classes $l2$ and $l3$ correspond to the events *early* and *late* peak [6] representing the alignment of f0 peaks and syllable nuclei.

9.3. Global Contours

The extracted global contours represent declination and inclination lines of different slopes. There are still open questions concerning the modelling of the global contours. First, a segmentation of a contour into intonational phrases guided by pitch discontinuities is not completely adequate since also boundary tones and prominent pitch accents correlate with such discontinuities. Here the syllable length between successive pitch discontinuities could help to distinguish between such prosodic events and real phrase boundaries, since the domain of pitch accents and boundary tones is in general limited to one syllable. The second open question concerns the amount of pitch reset. In this study we use the original phrase initial baseline values as starting points for the declination line. One potential approach is the prediction of pitch reset by a linear combination of factors like the final frequency of the preceeding intonation phrase, the durations of the preceeding and the current phrase, and their f0 slopes. A similar procedure was utilised to predict pause durations at prosodic boundaries in [3].

9.4. Generality of the Model

In this study we excluded other time-related prosodic aspects than intonation by time normalisation.

However, due to its data drivenness and generality our model is not just language independent but also principally capable to deal with other aspects of prosody than intonation. It would for example be of interest how it performs in modelling perceived local speech rate contours [9].

9.5. Relation to Linguistic Units

Another issue for future research is the question of linguistic significance of the extracted contour classes. They are only relevant for speech synthesis, if they can be related to linguistic dimensions like information and discourse structure. It is not yet known, whether the contour classes could be predicted from text.

10. References

[1] S. Rapp, “Automatisierte Erstellung von Korpora für die Prosodieforschung,” Ph.D. dissertation, University of Stuttgart, Institute of Natural Language Processing, Stuttgart, 1998.

[2] G. Möhler, “Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese,” Ph.D. dissertation, Institut für Maschinelle Sprachverarbeitung, Stuttgart, 1998.

[3] H. Pfitzinger and U. Reichel, “Text-based and Signal-based Prediction of Break Indices and Pause Durations,” in *Proc. Speech Prosody*, Dresden, 2006, pp. 133–136.

[4] M. Grice, M. Reyelt, R. Benzmüller, J. Mayer, and A. Batliner, “Consistency in Transcription and Labelling of German Intonation with GToBI,” in *Proc. ICSLP*, New Castle, Delaware, 1996, pp. 1716–1719.

[5] M. Reyelt, M. Grice, R. Benzmüller, J. Mayer, and A. Batliner, “Prosodische Etikettierung des Deutschen mit ToBI,” in *Natural Language and Speech Technology, Results of the third KONVENS conference*, D. Gibbon, Ed. Berlin, New York: Mouton de Gruyter, 1996, pp. 144–155.

[6] K. Kohler, “A model of German intonation,” in *AIPUK*, Kiel, 1991, vol. 25.

[7] G. Möhler and A. Conkie, “Parametric modeling of intonation using vector quantization,” in *Proc. 3rd ESCA Workshop on Speech Synthesis*, 1998.

[8] H. Fujisaki, “A note on physiological and physical basis for the phrase and the accent components in the voice fundamental frequency contour,” in *Vocal physiology: voice production, mechanisms, and functions*, O. Fujimura, Ed. New York: Raven, 1987, pp. 165–175.

[9] H. Pfitzinger, “Phonetische Analyse der Sprechgeschwindigkeit,” Ph.D. dissertation, Institute of Phonetics and Speech Processing, 2001.

[10] B. Möbius, *Ein quantitatives Modell der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläufen*. Tübingen: Niemeyer-Verlag, 1993.

[11] H. Mixdorff, “An Integrated Approach to Modeling German Prosody,” in *Studientexte zur Sprachkommunikation*. Dresden: Universitätsverlag, 2002, vol. 25.

[12] P. Taylor, “The rise/fall/connection model of intonation,” *Speech Communication*, vol. 15, pp. 169–186, 1995.

[13] T. Dunning, “Accurate methods for the statistics of surprise and coincidence,” *Computational Linguistics*, vol. 19, pp. 61–74, 1993.

[14] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken, “The MBROLA project: Towards a Set of High Quality Speech Synthesizers Free of Use for Non Commercial Purposes,” in *Proc. ICSLP*, vol. 3, Philadelphia, 1996, pp. 1393–1396.

[15] J. Nelder and R. Mead, “A Simplex Method for Function Minimization,” *Computer Journal*, vol. 7, pp. 308–313, 1965.

[16] D. Klatt, “Discrimination of fundamental frequency contours in synthetic speech: implications for models of speech perception,” *Journal of the Acoustical Society of America*, vol. 53, pp. 8–16, 1973.

[17] J. Van Santen, T. Mishra, and E. Klabbers, “Estimating Phrase Curves in the General Superpositional Intonation Model,” in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, 2004, pp. 61–66.

[18] A. Savitzky and M. Golay, “Smoothing and differentiation of data by simplified least squares procedures,” *Analytical Chemistry*, vol. 36, pp. 1627–1639, 1964.

[19] J. Pierrehumbert, “The phonology and phonetics of Englisch intonation,” Ph.D. dissertation, MIT, Cambridge, MA, 1980.

[20] ——, “Synthesizing intonation,” *Journal of the Acoustical Society of America*, vol. 70, no. 4, pp. 985–995, 1981.

[21] A. Black and A. Hunt, “Generating F0 contours from ToBI labels using linear regression,” in *Proc. ICSLP*, vol. 3, Philadelphia, 1996, pp. 1385–1388.

[22] J. Goldsmith, “Autosegmental Phonology,” Ph.D. dissertation, MIT, Cambridge, 1976.

[23] L. Adriaens, “Ein Modell deutscher Intonation: eine experimentell-phonetische Untersuchung nach den perzeptiv relevanten Grundfrequenzänderungen in vorgelesenem Text,” Ph.D. dissertation, University of Technology, Eindhoven, 1991.

[24] J. t’Hart, R. Collier, and A. Cohen, *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press, 1990.

[25] D. Klatt and L. Klatt, “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.