How useful are polynomials for analyzing intonation?

Laura E. de Ruiter
Max Planck Institute for Psycholinguistics
Nijmegen (Netherlands)

Laura.Herbst@mpi.nl

ABSTRACT

This paper presents the first application of polynomial modeling as a means for validating phonological pitch accent labels to German data. It is compared to traditional phonetic analysis (measuring minima, maxima, alignment). The traditional method fares better in classification, but results are comparable in statistical accent pair testing. Robustness tests show that pitch correction is necessary in both cases. The approaches are discussed in terms of their practicability, applicability to other domains of research and interpretability of their results.

1 Introduction

In intonation research, we are interested in the types of accents of a given language, or in a certain linguistic context. For many languages, descriptions of intonational grammars (e.g., in the auto-segmental metrical framework) have been proposed, and are widely used in the research community. The standard way of annotating intonation data is to label manually pitch tracks according to pre-specified labeling guidelines (e.g., ToBI, GToBI). However, it often remains unclear to what extent these labels are empirically valid. The standard procedure of multiple labelers can test their coding *reliability*, yet it cannot provide support for the labels' *validity*. The question is whether the assumed categories do actually constitute distinct classes. Can the theoretically postulated labels be clearly connected to measurable properties of the speech signal?

There might be more phonetic classes than phonological ones. We might find seven different distinct accent types, and that two or more of those are perceptually equivalent for speakers of that language. What should not be the case, however, is that researchers postulate accent classes for which we cannot find corroborative acoustic or perceptual evidence.

One way to validate labels is to take a number of phonetic measures (like f_0 excursion or alignment of peaks and valleys with the segmental string) and to test whether the assumed phonological categories differ from each other in these measures (e.g., [6]). Problems arise where pitch events like turning points cannot be unambiguously located because of microperturbations due to voiceless phonemes, for example. Here labelers have to make decisions, which are sometimes arbitrary, of where to locate the turning points.

A different approach is to model intonation contours mathematically, using polynomials (e.g., [1, 5, 7]). [7] used polynomials to describe entire intonational phrases, while in this study I am interested in individual pitch accents. Therefore, I follow the approach used by [5], where third order orthogonal polynomials (Fig.1) are used to model pitch accents.

Statistical analyses showed that the majority of the hand-labeled accent types differed significantly from each other in at least one coefficient. The authors conclude that polynomial modeling can provide intonational phonologists with a tool to empirically test linguistic descriptions of intonation (p.299).

In the present study, I use a corpus of German data to compare the two approaches. For reasons of brevity, I will refer to the first-mentioned approach as "LH" (low-high), and to the polynomial approach as "PN". I test how useful these methods are in validating phonological labels.

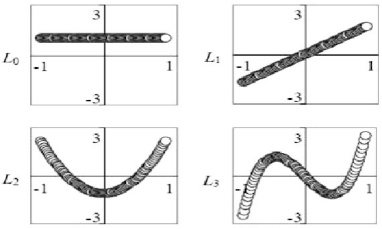

Figure 1: *Legendre polynomials L0-L3 (from [5], p. 288).*

At the same time, I evaluate their practicability, and test how robust they are when faced with only minimally pre-processed data. Thirdly, I discuss the two approaches in terms of the interpretability of their results. Finally, I consider their potential for intonation research in other domains of linguistic research. It would be desirable if these methods were also to provide us with a method to detect accent classes in production data more objectively and efficiently.

I first describe the procedure of determining the parameters for modeling f_0 for both methods (section 3). In section 4, I use these parameters in classification trees to predict phonological (GToBI) labels, and perform statistical analyses to see whether the parameters distinguish relevant dimensions for categorization. In section 5, I discuss the advantages and problems of both models in more detail.

2 Data

The data consisted of 135 utterances, spoken by 31 adult native speakers of German (6 male, 25 female). The recordings were made in an unechoic chamber, using a condenser microphone (Audio Technica AT4033A) and a DAT recorder (Tascam DA-45HR) at a sampling rate of 44 kHz (16 bit format).

About half of the utterances (63) were spontaneous productions from an elicitation task ([6]), the other half (72) were read-out sentences. The analyzed target words were all disyllabic, mainly sonorant words with stress on the first syllable (e.g., *Biene* (['bi:nə] 'bee'). I only chose words that occurred intonation-phrase (IP-) finally (presence of a boundary tone and a pause of at least 25 ms).

3 Representation of accents

3.1 Pre-processing and phonological annotation

The data were annotated and analyzed using Praat ([3]). For both methods, the IP was first segmented at the syllable level.

Accepted after peer review of full paper

For one condition (section 4.1.1), the pitch tracks were manually corrected for octave errors. To test the robustness of the two models (section 4.1.2), I also kept a set of uncorrected pitch files. In the next step, intonation of the target words was labeled following GToBI guidelines ([4]).

In order to be able to normalize for differences in f_0, I calculated each speaker's mean by taking the average f_0 of the first unstressed syllables of each utterance.

3.2 LH-method

The onset and offset of the stressed syllable (SS) were marked. Then, the absolute position and value of local f_0 maximum (max) and minimum (min) were determined manually. The domain in which these landmarks were identified consisted of the SS, the preceding syllable (PRE-S) and the syllable following it (POST-S). Note that for H+!H*, the max was taken to be the high on PRE-S, while min was set at the peak in the SS, if there was a clear 'bump'. If not, the middle of the SS was marked. I then normalized all f_0 values by dividing them by the speaker mean. The f_0 excursion was calculated as the absolute difference between the minimum and the maximum. Alignment values were calculated relative to the duration of the SS. The parameters that were used to describe a given intonation contour were:

- Normalized f_0 maximum (NORMMAX)
- Normalized f_0 minimum (NORMMIN)
- Excursion (absolute value of norm. f_0 maximum – norm. f_0 minimum, in Hz, EXC)
- Relative position of f_0 maximum (POSMAX)
- Relative position of f_0 minimum (POSMIN)

3.3 PN-method

To ensure comparability, the procedure was kept similar to the one described in [5], but simplified in certain aspects, in particular weighting. Like in [5], my model was specified by a set of coefficients, c_i, that multiply the different Legendre polynomials before they are added together:

$$M(x) = c_0 + c_1 \cdot x + c_2 \cdot (\frac{1}{2}(3x^2 - 1)) + c_3 \cdot (\frac{1}{2}(5x^3 - 3x))$$

F_0 was measured in steps of 5ms. Before fitting the data, I applied two normalization procedures: All f_0 values were divided by the speaker mean f_0, and the time axis of the analysis domain (voiced region, see specifications below) was shifted and scaled to values between -1 and 1, which is a prerequisite for modeling using Legendre polynomials.

For the estimation of the coefficients of the Legendre polynomials that best describe a given intonation contour, I used Polyfit, a customized computer program written in C++ ([8]). The program reads in normalized f_0 values and a weighting parameter (described below) and calculates those Legendre coefficients that minimize the difference between the predicted polynomial and the original pitch contour as estimated by Praat's pitch tracking algorithm. The quantity that is minimized is a chi-square related merit function:

$$m = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{data_i - pred_i}{w_i}\right]^2$$

where w_i is a weighting quantity indicating the relative contribution of data point i to the merit function. The program uses a well-known General Linear Least Squares algorithm based on normal equations and Gauss-Jordan elimination, and is described in detail in [9].

The weighting parameter w combines intensity and a periodicity measure (harmonics-to-noise ratio, HNR) to give more weight to loud and sonorant regions. A higher w for a certain time window forces the algorithm to model f_0 values in this region with more precision. I determined intensity and HNR (in dB) at each point at which f_0 was measured. Intensity was normalized by dividing each value by the mean intensity of the voiced parts of the entire utterance. Unlike intensity, HNR values usually cover a wider range of values and can also be negative, in cases where there is more noise than harmonics in the signal. I normalized the HNR measures using a sigmoid function, which transforms all possible values (from ∞ to -∞) into values from 0 to 1. Hence negative HNR values receive a low score near 0, whereas positive ones receive a score closer to 1. I calculated the parameter k using the criterion that a value of 15dB (roughly equal to 97% energy from the harmonic part, cf. Praat manual on "harmonicity") receives an H-score of 0.75. The resulting coefficient k is 0.02453.

$$H(t) = \frac{1}{1 + e^{-kt}}$$

The weighting parameter w used in the fitting program was the product of normalized intensity (ι) and the standardized HNR value (H):

$$w = \iota \cdot H$$

As with the LH-method, the domain for pitch measurements consisted of three syllables: PRE-S, SS and POST-S. However, unvoiced regions (like devoiced vowels) at the beginning or the end of the domain can be problematic. When the program determines the coefficients to model the intonation contour, it mainly fits the polynomials to the voiced parts while the polynomials can take any form for unvoiced regions. This is not harmful for voiceless regions in the middle of a voiced region, assuming that the f_0 contour constitutes a smooth function. However, for voiceless regions before pitch onset or after pitch offset, the fitting becomes unpredictable. Note that weighting alone cannot solve this problem, as a very low w would still 'allow' the program to fit almost any curve. To avoid this problem, I set the domain to start at the first voiced frame within the original three-syllable domain, and to end at the last one.

4 Analysis and results

4.1 Classification

4.1.1 Corrected pitch

From both methods I obtained a data set with 135 data points and four parameters each: EXC, POSMAX, POSMIN and NORMMAX in the case of the LH-model, and the coefficients of the first four Legendre polynomials, L0, L1, L2 and L3 for PN-model. For reasons of clarity, I will refer to the four coefficients as AVERAGE, SLOPE, PARABOLA and WAVE, following the naming convention used by [5].

I then investigated to what extent the class of a given data point (i.e., its phonological label) can be predicted from those parameters. Classification trees were built using the Recursive Partitioning and Regression Trees function in R ([10]) to predict GToBI labels (Fig. 2).

The resulting trees for both methods look very similar. Deaccentuation, H+!H* and H* cluster together in the left major branches of the tree, while the rising accents L* and L*+H are found in the right-hand branch. The classification algorithm that uses the LH data needed three parameters (EXC, POSMAX, and POSMIN), for the PN-based algorithm two parameters (SLOPE and PARABOLA) suffice.

The accent type H+L* cannot be predicted from the PN data set, but from the LH set, and two other accent types (which occur less than 7% of the cases) do not show up as terminal leaves in either tree (!H* and L+H*). Overall, the LH-based algorithm is more successful in predicting GToBI labels: On average, 71% of a given accent type was correctly classified by the LH-model, as compared to 62% for the PN-model. The mean classification error for the LH-model was 38%, while it was 47% for the PN-model. (This is because GToBI descriptions are based on the LH-type analysis, see Discussion.)

4.1.2 Uncorrected pitch

I also tested how both models performed in classifying accents when they had to deal with uncorrected pitch. The same procedure was applied as before, but this time global min and max were marked automatically in the voiced region for the LH-analysis. This was to test the model in the worst-case scenario (uncorrected pitch, automatic annotation).

In the case of LH, the number of leaves was reduced from six to four; the accent types H+L* and L* could no longer be predicted. In contrast, the PN-model-based tree had now more terminal leaves (7) than in the pitch-corrected case (5), which was due to now three terminal nodes that predicted deaccentuation (instead of one). The PN-model's overall hit rate in classifying a given accent decreased from 64 to 53%, while the LH-model's rate improved slightly from 71 to 73%. It should be borne in mind that fewer leaves increased the possibility of classifying an accent correctly merely by chance (25%).

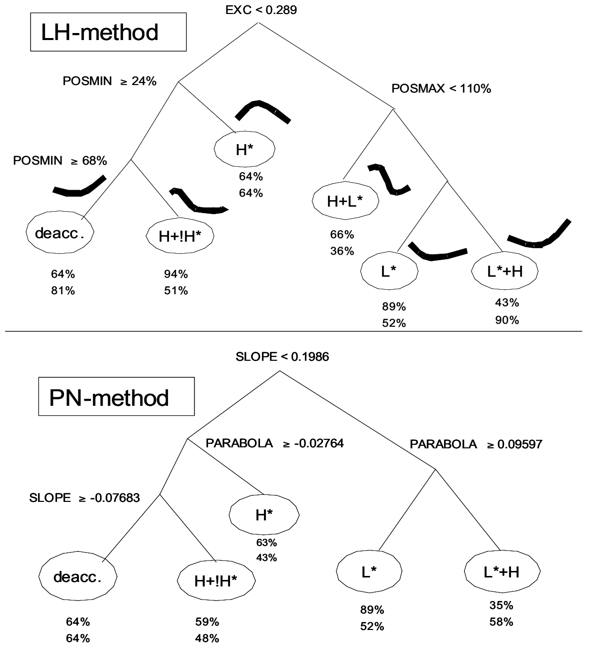

Figure 2: *Classification trees (pruned) for both methods, with stylized accent shapes. Numbers indicate the overall % of correct classification of the accent (first row) and the purity of the leaf (i.e., the proportion of that accent type; second row).*

4.2 Accent pair contrasts

I also carried out statistical analyses with R ([10]) comparing all accents with each other, similar to [5]. Linear mixed effects (LME) models were built, using subject and word as crossed random factors (where applicable), and the three parameters that turned out to be relevant each as fixed factors. These were EXC, POSMAX and POSMIN in the LH-model, and SLOPE, PARABOLA and WAVE in the PN-model. The associated p-values were obtained by Markov chain Monte Carlo sampling and adjusted using Holm's correction for multiple comparisons.

Figure 3 gives an overview of the differences. No differences were found between H* and L*, and between H* and H+L* in the PN-model. Neither model found any differences between L*, L*+H and L*+H. The PN-model found a significant difference between H* and L+H*, while this was not the case in the LH-model.

LH-model

PN-model

	deacc. (N = 34)	H* (N = 25)	L+H* (N = 9)	L* (N = 19)	L*+H (N = 20)	H+L* (N = 9)	H+!H* (N = 17)
deacc. (N = 34)		■	▒	■	■	■	■
H* (N = 25)	▨			▒	▒	■	■
L+H* (N = 9)	▨	☰		▒		■	■
L* (N = 19)	▨		▨			■	■
L*+H (N = 20)	▨	▨				■	■
H+L* (N = 9)	▨		▨	▨	☰		
H+!H* (N = 17)		☰	▨	▨	▨		

Figure 3: *Matrix of differences found between accent types using LME-models. Black/striped indicates p<0.01, gray/gray-striped indicates p<0.05.*

5 Discussion

Both methods delivered comparable results and appear suitable to empirically test linguistic descriptions of intonation, that is, accent type labels. By and large the models picked up most of the differences one would expect.

It is not surprising that the LH-approach fared better in predicting the GToBI labels, as the data were labeled using LH-model parameters implicitly as criteria. Still, the classification shows that the PN-method, too, yields data that can be used to automatically form sensible groupings of accents.[1] The finding is interesting given that the PN-model worked with less 'linguistic' information. Features like alignment (here operationalised as POSMIN, POSMAX) seem to be determined phonologically (see e.g., [2]), yet the lack of such information did not seem to impede the performance of the PN-model dramatically. The model's failure to distinguish L* and H* in the accent contrasts remains puzzling, but this is not caused by the absence of alignment information. Another unexpected finding is that the LH-model did not find any differences between L*, L*+H and L+H*, three accent types for which POSMIN should be the discriminating feature [4].

I will now discuss the other aspects mentioned in the introduction in turn.

[1] Note that unlike [5], I also included deaccentuation in my data set. The results show that it is possible to characterize deaccentuation in terms of the PN-model

5.1 Robustness

Both models' success in predicting phonological labels suffered from uncorrected pitch, though in different ways: The LH-model was not able to predict more than four accent types, while the PN-model's overall accuracy was reduced. It should be pointed out that for the LH-method, I compared the best-case scenario (hand-corrected pitch, manually labeled min/max) with the worst-case scenario (uncorrected pitch, automatic detection of min/max). Intermediate solutions (e.g., corrected pitch, manually set min/max) may offer an acceptable trade-off between time invested and accuracy. For the PN-model, the use of corrected pitch tracks seems advisable.

5.2 Practicability

The LH-approach is straightforward and easy to apply. Pitch values and timing information for maxima and minima need to be extracted. However, the hand-correction of pitch and turning points is a time-consuming process, which often seems necessary (see 5.1). In the case of turning points in particular this process is not only slow, but also problematic, in that the labelers tacitly smooth out the pitch curve, thereby making the exact location of the minimum or maximum uncertain. Yet in my study it was above all the alignment of those points that was critical in the LH-approach for categorizing accents.

This tedious task is not necessary in the PN-approach. Here, f_0 irregularities are smoothed out by the curve-fitting algorithm. Still, even though microperturbations have not to be taken into consideration, correcting octave errors improves classification success considerably, so that this step also seems necessary in this approach. Furthermore, the PN-method requires a customized program and scripts to calculate the weighting parameter, resources which may not be accessible to everyone.

On the whole, the LH-approach takes more time for preprocessing of the data, while the PN-approach is more costly in terms of "tool development" time. Once these tools are up and running, however, the PN-method can swiftly be applied to larger amounts of data. It remains a problem for both approaches that the same phonological category is often realized differently depending on the (phonetic) context. Therefore some a priori decisions, like the choice of the analysis domain, will have to be taken.

5.3 Applicability

Both methods can help validate postulated categories. However, not always are the underlying categories known. This is the case for example with learner' speech (both first and second language acquisition), or prosodically undescribed languages. The LH-approach may prove less useful here, as the acoustic measurements and segmental landmarks important for the prosodic system in question are largely unknown. One would have to label (and correct) a larger number of parameters and see to what extent these can be used to form sensible groups of accents (e.g., by clustering). These choices need not be made when polynomial modeling is used. The dimensions that describe a contour mathematically are defined by the model. At the same time, potentially important linguistic information gets lost. However, in this study at least, this did not seem to be a problem.

Another domain of application for the data is speech synthesis, both for commercial purposes and for the use in linguistic perception experiments. Here the PN-approach has an advantage: An intonation curve can be fully reconstructed from the polynomial function. In order to do this from LH-data, one needs at least five measurements: POSMIN, POSMAX, EXC, NORMMAX (or NORMMIN), and pitch offset.

5.4 Interpretability

LH-parameters like excursion and relative position of min and max are easily visualized and understood. As explained in [5] (p. 289), the first coefficients of the PN-model can be linked directly to physical properties of the (pitch) curve. However, these are not expressed in the units well known to linguists, which may make their interpretation initially more difficult.

6 Conclusion

In this paper I have applied two different methods of validating phonological pitch accent type labels, labeling minima and maxima, as well as third order polynomials.

Both in classification and in the pair-wise comparison, the standard model fares better compared to the polynomial model. From this point of view, the LH method is more useful to the researcher who wants to check that a given set of labeling criteria was applied appropriately. On the other hand, I have shown that the polynomial model avoids some of the pitfalls that come with the 'traditional' analysis, such as the location of minima and maxima. Polynomial fitting also seems to be more efficient when larger amounts of data are analyzed, because it relies less on annotation by hand. It may therefore also be useful for explorative analyses of prosodic data.

The PN-method has only recently been introduced to the field of intonational phonology. My preliminary conclusion is that it can be a useful tool in this area, but more research will be needed to put it to the test. Future versions may increase sophistication by incorporating linguistic information and may turn out to be more powerful than the traditional method.

7 Acknowledgments

I would like to thank Jan Peter de Ruiter for writing the Polyfit program, and Bettina Braun for help with this paper.

8 References

[1] Andruski J., and Costello, J. (2004). *Using polynomial equations to model pitch contour shape in lexical tones: An example from Green Mong*. J of the Intern. Phonetic Assoc. (34), 125-140.

[2] Arvaniti, A., Ladd, D. R., and Mennen, I. (2006). *Tonal association and tonal alignment: Evidence from Greek polar questions and contrastive statements*. Lang Speech 49(4), 421-450.

[3] Boersma, P., and Weenink, D. (2008). *Praat: doing phonetics by computer* (version 5.0.18) [computer program].

[4] Benzmüller, R. et al. *Trainingsmaterialien zur Etikettierung deutscher Intonation mit GToBI - Version2*.

[5] Grabe, E., Kochanski, G., and Coleman, J. (2007). *Connecting intonational labels to mathematical descriptions of fundamental frequency*. Lang Speech 50(3), 281-310.

[6] Herbst, L. E. (2007). *German 5-year-olds' intonational marking of information status*. Proc. of the 16th ICPhS 2007, 1557-1560.

[7] Hirst, D.J., di Cristo, A., and Espesser, R. (2000). Levels of representation and levels of analysis for intonation. In: M. Horne (ed) *Prosody: Theory and Experiment*. Dordrecht: Kluwer.

[8] De Ruiter, J.P. (2008). *Polyfit 1.0*, fitting Nth order orthogonal polynomials on weighted datasets [Computer program] Available on request from janpeter.deruiter@mpi.nl.

[9] Press, W. H. et al. (1988). *Numerical Recipes in C; The Art of Scientific Computing*. Cambridge: Cambridge University Press.

[10] R Development Core Team (2005). *R: A language and environment for statistical computing* (v. 2.6.2) [computer program].