Empirical Validation of Hand-labelled Nuclear Accent Patterns

E. Grabe, G. Kochanski & J. Coleman

Phonetics Laboratory
University of Oxford, United Kingdom
{esther.grabe; greg.kochanski; john.coleman}@phon.ox.ac.uk

Abstract

In a corpus containing speech data from seven dialects of English, we hand-labelled over 700 nuclear accents and identified seven accent types. Then we used four-term mathematical models to describe the fundamental frequency patterns associated with the accents. A statistical analysis showed that the models for six of the seven accents differed significantly from each other. Our hand-labels were associated with consistently different f_0 patterns.

1. Introduction

Mathematical models of intonation used in speech technology are often inaccessible to linguists. By the same token, phonological descriptions of intonation are rarely used by speech technologists, as they cannot be implemented in software. In this paper, we explore bridges between intonational phonology and speech technology. Phonologists need methods that allow for empirical validation of labelling systems and access to larger bodies of data. Speech technologists require empirically tested and directly implementable models filtered by linguistic insights.

A first step in this direction was taken by Andruski and Costello [1]. Andruski and Costello used coefficients from polynomial equations to explore small differences in the f_0 contours of three low falling tones in Green Mong. Polynomial equations are a common mathematical approach to the description of curves; they produce a hierarchy of descriptions of increasing complexity and accuracy. Mathematically, they are expressions involving a sum of powers in one or more variables multiplied by constants (e.g. $a_2x^2 + a_1x + a_0$). In work on intonation in speech technology, polynomial equations constitute one of several standard approaches to curve-fitting. Other well-known curve-fitting models of intonation are described in Fujisaki [2], Hirst, di Christo and Espesser [3] and Taylor [4].

In their investigation of Green Mong, Andruski and Costello [1] used polynomial equations to test whether f_0 contour shape alone could distinguish the three falling tones. They estimated linear and quadratic equations for each pitch contour ($y = a + bx$ and quadratic $y = a + bx + cx^2$, respectively). The resulting coefficients (a, b, c) provided a quantitative description of the slope and the curvature of the three tones. Subsequent analyses revealed that the three tones could indeed be discriminated well above chance level on the basis of contour shape.

In the present paper, we use polynomial equations to describe the rich inventory of nuclear accents found in English spoken in the British Isles [5]. We show how autosegmental-metrical accent labels can be mapped onto relatively simple polynomial models to provide quantitative, statistically testable descriptions of each accent type.

2. Method

Our research was based on 714 read sentences in the IViE corpus [5, 6]. These were produced by three male and three female speakers from each of seven dialects of English spoken in London, Cambridge, Leeds, Bradford, Newcastle, Belfast, and Dublin. The London speakers were of West Indian descent and the speakers from Bradford were English-Punjabi bilinguals. The sentences consisted of fully voiced declaratives, *wh*-questions, polar questions and declarative questions, read in isolation. They are listed on our web-site [6] and in [7].

2.1. Autosegmental-metrical intonation labels

We assigned autosegmental-metrical intonation labels to the 714 sentences via a combination of auditory analysis and visual inspection of fundamental frequency traces, a standard approach in the field [8, 9]. We used the IViE system, an autosegmental-metrical intonation transcription system developed for labelling of dialectal intonational variation in English [5, 10]. Transcriptions were made using the ESPS/xwaves+ package developed by Entropic Research Laboratories. A completed transcription consisted of an audio file, a time-aligned fundamental frequency trace and time-aligned text files containing transcriptions of intonation patterns. The labeling procedure is described in [5].

Seven nuclear accent types were labelled in more than five instances and were included in the present study: H* H% (high rise), H*L % (fall), H*L H% (fall-rise), L*H L% (rise-(plateau)-fall), L*H H% (rise), L*H % (rise-plateau) and L* H% (late rise). (The '%' boundary symbol indicates that the f_0 level associated with the last tone of the last accent in the intonation phrase is continued up to the boundary.) The number of tokens of each nuclear accent type in the data set is shown in Table 1.

Table 1: *Distribution of nuclear accents in the sentence data in the IViE corpus.*

Nuclear accents		Tokens
H*L %	fall	414
L*H %	rise-plateau	187
H*L H%	fall-rise	41
L*H H%	rise	32
H* H%	high rise	15
L* H%	late rise	12
L*H L%	rise-plateau-fall	9
		710

Table 1 shows that the frequency distribution of nuclear accent types was uneven, as one would expect in a large speech corpus ('lopsided sparsity', van Santen [11]). In the present study, we handle data sparsity via Multivariate Analyses of Variance (MANOVA), a statistical technique developed to process uneven amounts of data.

2.2. Mathematical modelling

A detailed description of our approach to polynomial modelling, including instructions for how to carry out modelling in MS Excel, is given in [12]. Our analysis was carried out with a set of custom-written Python scripts. A brief description follows.

We used Legendre polynomials. These are *orthogonal*, consequently, there are no correlations among the coefficients that describe the shape of an intonation pattern. Each nuclear accent was modelled separately. The analysed region of f_0 began 100 milliseconds before the nuclear accent of the sentence (as defined by the final accent label preceding a boundary), and extended to the end of the voiced part of the sentence. The central step in the analysis was to represent the data as a best-fit sum of Legendre polynomials where each polynomial is normalised to have unit variance. The result of the analysis was a model for the f_0 contour of each accent. The model was specified by a set of coefficients, c_i, that multiply the different Legendre polynomials (L_i) before they are added together:

$$M(x) = \sum_i c_i \cdot L_i(x) \qquad (1)$$

Next, a set of coefficients was found that gave the best fit of Equation 1 to the data. To find these, we used a weighted linear maximum-likelihood regression, exactly as in [13]. We limited our models to four coefficients. These described the average and the slope of each f_0 contour in the data, and two kinds of curvature: a parabola shape and a wave shape.

3. Results

We carried out an Analysis of Variance to test whether the polynomial models associated with each of the seven accents were statistically different. (Note that all results given in this section apply to the data set as a whole, not on a per-dialect basis. The size of the differences is discussed in section 4).

The dependent variables were AVERAGE (c_0), SLOPE (c_1), PARABOLA (c_2) and WAVE (c_3). LABEL (i.e. nuclear accent type) was the independent variable (Table 1). The analysis produced very highly significant main effects of LABEL on the dependent variables (AVERAGE $F[1, 6] = 54.0$, $p < 0.001$, SLOPE $F[1, 6] = 78.6$, $p < 0.001$, PARABOLA $F[1, 6] = 14.4$, $p < 0.001$, WAVE $F[1, 6] = 15.2$, $p < 0.001$).

Post-hoc Tukey tests showed that 17 of the 21 accent pairs differend significantly at $p < 0.001$, in one or more coefficients. A further two pairs differed at at $p < 0.05$. We did not find significant differences between L* H% (a late rise observed in data from London) and the other two low rising accents L*H % (the rise plateau, common in Belfast English) and L*H H% (the rise, observed in all dialects).

Finally, the statistical analysis showed that a model based on three coefficients would also have been successful in distinguishing between the nuclear accent contours in our corpus. We found significant differences between contours in the fourth coefficient, but the information was redundant.

Figure 1 shows mean coefficient values for each nuclear accent type (but recall that accents were modeled individually).

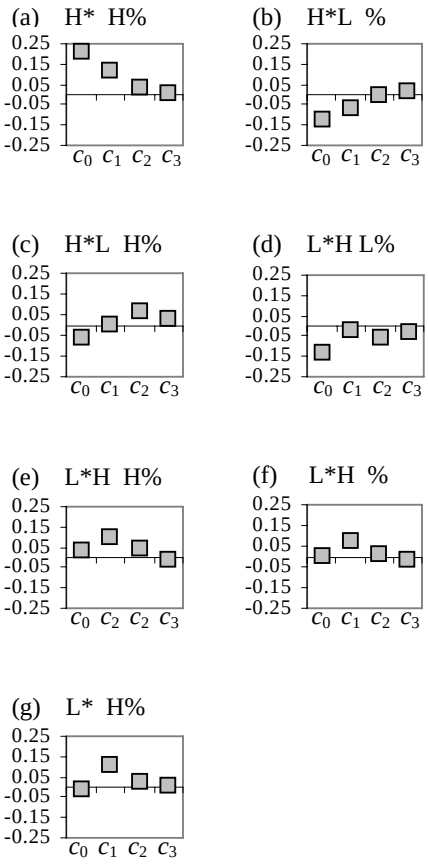

Figure 1: Four-coefficient f_0 profiles for seven nuclear accents in the IViE corpus. The coefficients are listed on the x-axis. The y-axis shows units of normalised f_0 (0.1 = 10% of the speaker's average f_0).

In Figure 1, a negative coefficient c_0 is equivalent to a low average f_0 for the accent type; a positive coefficient c_0 shows the opposite. A negative coefficient c_1 shows that the accent has falling slope, a positive c_1 represents a rising slope. A negative coefficient c_2 models a cup-shape, a positive c_2 describes a domed shape. A negative coefficient c_3, shown for completeness, describes a falling-rising-falling component of the shape and a positive c_3 describes a rising-falling-rising component.

We will now describe two of the profiles shown in Figure 1, by way of example: Figure 1a shows the four-coefficient representation of H* H%, the high rise. The first coefficient was positive and large: H* H% accents had a relatively high average. The second coefficient was also positive and large: H* H% accents had rising slopes. The third coefficient was small but positive: the pattern was somewhat cup-shaped. The fourth coefficient was close to zero, indicating that the WAVE component contributed little to the shape.

Figure 1e shows L*H H%, a rise from a low accented syllable. The first coefficient, the average, was much lower than for H* H%; the accent began significantly lower in the speakers' f_0 ranges. The second coefficient was large, as expected for a rising slope. The small but positive third coefficient shows that L*H H% accents were also somewhat cup-shaped. Again, the average of the fourth coefficient was about zero.

To illustrate further the plausibility of the orthogonal polynomial descriptions, in Figure 2 we show an f_0 model for each accent shape, reconstructed from the four coefficients in Figure 1.

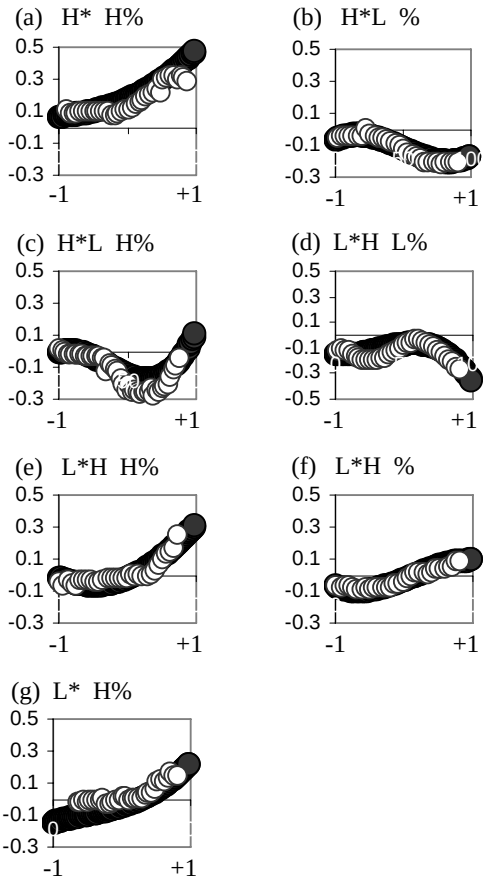

Figure 2: Reconstruction of f_0 models from the four coefficients (thick black lines) with superimposed f_0 data (unfilled circles). The x-axis shows normalised time (-1 = beginning of utterance; +1 = end of utterance). The y-axis shows normalised f_0.

In Figure 2, the reconstructed f_0 models (thick black lines) summarise the salient characteristics of each accent type. The reconstruction was done by entering the relevant set of coefficients into Equation 1 and computing $M(x)$ for 100 different values of x between –1 and 1. For comparison, we have superimposed one original, normalised f_0 trace from the IViE corpus (unfilled circles) in each panel. This superimposed trace has the least mean-square difference from the model. The traces show that the polynomial models – despite being an average – are representative of the data.

4. Discussion

The models in Figures 1 and 2 provide a quantitative link between autosegmental-metrical intonation labels and statistical characteristics of classified accents. The figure shows that each label is associated with a different contour. With the exception of L* H%, the late rise, all contours are statistically different.

The f_0 traces labelled as L* H% could not be distinguished significantly from those labelled as L*H H% or those labelled as L*H %. One might conclude that L* H% is not a separate accent type and the label should be collapsed with another label describing a rise. The conclusion is not, however, straightforward: firstly, the results of the statistical analysis do not show which accent L* H% should be collapsed with, L*H H% or L*H H%. Secondly, since we worked with very few tokens of L* H%, we cannot entirely dismiss the issue of data sparsity (cf. Table 1). MANOVA looks for statistical differences between the means of the distributions of coefficients associated with different labels. Given more data, the approach becomes more sensitive: means become more precisely defined as more measurements are made. Had we worked with a larger number of L* H% accents, a significant difference might have emerged. This argument points out a limitation of a purely statistical analysis: any difference between the coefficients of two labels, however small, could be statistically significant if one had a large enough corpus. Statistical significance is only meaningful if coupled with an estimate of the size of the effect. In our data, in addition to being statistically significant, some of the differences are quite large and should be perceptually obvious. H*L % and H*L H%, for instance, are not the most different pair but differ by 0.2 normalised f_0 units at the end of the utterance. For a speaker with a 170 Hz mean f_0, this would be a difference of 34 Hz, substantially larger than segmental perturbations and the psychophysical just-noticeable-difference.

Finally, in our examination of the result for the late rise L* H%, we need to consider the effect of neutralisation. The shape of a nuclear f_0 contour is affected by the structure and number of syllables available. In British English, distinctions between L*H H%, L*H % and L* H% can be observed only if the accented syllable is followed by at least one syllable; otherwise, patterns are compressed or truncated [14]. If the accented syllable is followed by two or more syllables, differences become obvious. In our materials, nuclear accents were produced on disyllabic trochees. Had our rises been produced on longer words, the distinctions between all of them might have been statistically significant.

5. Conclusions

Our approach shows that intonational phonological hand-labels can be supported by empirical acoustic evidence. We found that six out of seven impressionistically assigned labels were associated with a set of statistically different f_0 patterns.

Our methodology has a number of applications. Firstly, and most obviously, linguists can use the approach to investigate empirically the acoustic basis of their intonational phonological classifications. Secondly, at least potentially, the approach may provide linguists with access to larger bodies of data. In collaboration with speech technologists, intonational phonologists could develop methods that allow for automatic

classification of large numbers of accents. Data from large corpora would allow for descriptions of accent usage in different texts and styles and by different speakers. Moreover, with large corpora, rare accent patterns could be detected.

The approach can also add to work on the alignment of intonation with segmental anchors, that is, vowels, consonants and syllable boundaries [15, 16, 17, 18, 19, 20]. Polynomial models of f_0 can capture changes in the average, slope and curvature of a contour and this information can usefully supplement (or in some cases, replace) hand-measurements. A stylised example illustrating how polynomial modelling can contribute to work on alignment is given in Appendix C in [12].

More generally, the approach allows for a combination of qualitative and quantitative comparisons of intonation systems across dialects and languages. Cross-linguistic and cross-dialectal differences may involve the phonology or the phonetics of intonation or a combination of both. A combined qualitative/quantitative approach to analysis can provide new insights.

Finally, the models are of value to speech technologists. Since the models are based on insights from linguistics, they are, in a sense, pre-filtered. Hand-labellers have determined the existence of an accent and the location of the stressed syllable, and they have decided on the equivalence of patterns on texts with different distributions of voicing and different numbers of syllables. But unlike hand-labels, the 'translated' data can be implemented directly in a synthesis or recognition system.

We conclude that polynomial modelling is of value to intonational phonologists and may help to fill the gap between intonational phonology and speech technology. Our results have shown that impressionistically salient aspects of f_0 in nuclear accents can be expressed quantitatively, using a small number of mathematical terms. This approach allows for empirical testing of linguistic descriptions of intonation and opens up new avenues for collaboration.

6. Acknowledgements

This research was supported by a grant from the UK Economic and Social Research Council to Esther Grabe and John Coleman (RES000–23–0149).

7. References

[1] Andruski, J.; Costello, J., 2004. Using polynomial equations to model pitch contour shape in lexical tones: an example from Green Mong. *Journal of the International Phonetic Association*, 34, 125 – 140.

[2] Fujisaki, H., 1992. Modelling the Process of Fundamental Frequency Contour Generation. In *Speech Perception, Production and Linguistic Structure*, Y. Tohkura, E. Vatikiotis-Bateson, Y. Sagisasaka (eds). Amsterdam: IOS Press, 3 – 328.

[3] Hirst, D.J., di Christo, A. & Espesser, R., 1993. Levels of representation and levels of analysis for the description of intonation systems. In *Prosody: Theory and Experiment*, M. Horne (ed.). Dordrecht: Kluwer Academic Publishers, 51 – 88.

[4] Taylor, P., 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107, 1697 – 1714.

[5] Grabe, E., 2004. Intonational variation in English. In *Regional Variation in Intonation*, P. Gilles; J. Peters (eds.). Tübingen: Niemeyer, 9 – 31.

[6] Intonation in the British Isles. Research project web-site. www.phon.ox.ac.uk/~esther/ivyweb/

[7] Grabe, E., Kochanski, G. & Coleman, J., 2005. The intonation of native accent varieties in the British Isles - potential for miscommunication? In *English pronunciation models: a changing scene*, K. Dziubalska-Kolaczyk; J. Przedlacka (eds.) Linguistic Insights Series: Peter Lang, 311 – 338.

[8] Ladd, D. R., 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

[9] Beckman, M. E.; Ayers Elam, G., 1997. *Guidelines for ToBI Labelling, version 3*. The Ohio State University Research Foundation, Ohio State University.

[10] Grabe, E. 2001., The IViE labelling guide web-site. Phonetics Laboratory, University of Oxford. http://www.phon.ox.ac.uk/~esther/ivyweb/guide.html

[11] van Santen, J., 1994. Using statistics in text-to-speech system construction. In *Proceedings of the ESCA/IEEE workshop on speech synthesis*, pp. 240 – 243, New Paltz.

[12] Grabe, E.; Kochanski, G.; Coleman, J. (submitted). Connecting Intonation labels to mathematical models of fundamental frequency. *Language and Speech*.

[13] Kochanski, G.; Grabe, E.; Coleman, J.; Rosner, B., 2005. Loudness predicts prominence; fundamental frequency lends little. *The Journal of the Acoustical Society of America 118(2)*, 1038-1054.

[14] Grabe, E., 1998. Pitch accent realisation in English and German. *Journal of Phonetics* 26, 129-144.

[15] Silverman, K.; Pierrehumbert, J., 1990. The timing of prenuclear high accents in English. In *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, J. Kingston; M.E. Beckman (eds.). Cambridge: Cambridge University Press. 72 – 106

[16] Prieto, P., van Santen, J., & Hirschberg, J., 1995. Tonal alignment patterns in Spanish. *Journal of Phonetics*, 23, 429 – 451.

[17] Arvaniti, A., Ladd, D. R., & Mennen, A., 1998. Stability and alignment of pitch targets in Modern Greek prenuclear accents, *Journal of Phonetics*, 26, 3 – 5.

[18] Ladd, D.R., Mennen, I., & Schepman, A., 2000. Phonological conditioning of peak alignment in rising accents in Dutch. *The Journal of the Acoustical Society of America*, 107, 2685 – 2696.

[19] d'Imperio, M., 2001. Tonal structure and pitch targets in Italian focus constituents, *Speech Communication*, 33, 339 – 356.

[20] Frota, S., 2002. Tonal association and target alignment in European Portuguese nuclear falls. In *Laboratory Phonology 7*; C. Gussenhoven; N. Warner (eds.). Berlin: Mouton de Gruyter, 387-418