Connecting Intonation Labels to Mathematical Descriptions of Fundamental Frequency

Esther Grabe, Greg Kochanski, John Coleman

University of Oxford

Key words

intonational phonology

polynomials

quantitative modeling

Abstract

The 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 directly in applications. Consequently, these research communities do not benefit much from each other's insights. In this paper, we explore the interface between the disciplines, in search of bridges between intonational phonology and speech technology.

In a corpus of speech data from seven dialects of English, we hand-labeled over 700 sentences and identified seven nuclear accent types. Then we fitted a third-order polynomial to the fundamental frequency (F0) contour in the region around the accent mark. The polynomial captures the local shape (time-dependence) of F0 in a few numbers, in our case, four coefficients. The coefficients were subjected to statistical analysis. Nineteen of the 21 pairs of accent types differed significantly in one or more coefficients. Our approach bridges the gap between intonational phonology and speech technology. It provides quantitative, empirically testable models of intonation labels that can be implemented in applications.

1 Introduction

Speech technology researchers are keen to add information carried by intonation to speech synthesis and recognition systems. Commercial speech synthesis systems cannot, for instance, produce the intonation patterns of different dialects. In speech recognition, the use of intonation is minimal, although the role of intonation in human speech comprehension is well established (for an overview, see Cutler, Dahan, & Donselaar, 1997). At first sight, the lack of success in exploiting much of intonation may be surprising; after all, in speech technology, there is no lack of quantitative data on the acoustic structure of intonation. The interpretation of this wealth of information is not, however, straightforward. Acoustic data on intonation must be

Acknowledgments: This research was supported by a grant from the U.K. Economic and Social Research Council to E. Grabe and J. Coleman (RES000-23-0149). For comments and suggestions, we are grateful to Burton Rosner, Tim Bunnell, and Steve Hoskins.

Address for correspondence. John Coleman, Phonetics Laboratory, 41 Wellington Square, Oxford University, OX1 2JF, U.K.

linked to the linguistic structure of the text: a nontrivial task, as there is no one-to-one mapping of particular intonation patterns onto particular meanings. Also, F0 contours associated with particular intonation patterns change with the distribution of voicing in the text and the number of syllables. Information from linguistic analyses of intonation such as ToBI (Beckman & Ayers Elam, 1997) is generally welcome but cannot be implemented directly.

Intonational phonologists have a comparable problem, but in reverse. Most cannot or do not access or interpret the data from speech technology, that is, large speech corpora. Instead, they work with specifically designed, phonetically controlled sentences or small corpora and describe the intonation patterns in the data by hand with impressionistic labels. Usually, these are autosegmental-metrical labels, consisting of combinations of two tones, H (high) and L (low) (cf. Ladd, 1996). The labels provide a classification of observed intonation patterns into combinations of a limited set of categorical choices. The underlying, still largely unproven assumption is that each categorical choice contributes to the meaning of an utterance. This contribution interacts with the text and the context and consequently, it is not easy to pin down (Pierrehumbert & Hirschberg, 1990). The process of assigning autosegmental-metrical labels to a corpus requires much effort from highly trained specialists. To produce reasonably reproducible sets of labels, labelers need a solid understanding of intonation theory (and preferably more than one theory) and speech acoustics, and they need to have received good ear-training. Consequently, labeled corpora are usually small. Nevertheless, as a technique, some form of hand-labeling is indispensable to the intonational phonologist. The process contributes significantly towards an understanding of the data and few phoneticians or phonologists would doubt that intonational phonology has produced many valuable insights (an overview is given in Ladd, 1996). The lack of a quantitative basis for labels in corpora, however, inevitably raises questions about their empirical validity.

To a certain extent, the validity of labeling systems for intonation can be tested with multi-labeler exercises (e.g., Grice, Reyelt, Benzmüller, Jun, Lee, Kim, & Lee, 2000; Mayer, & Batliner, 1996; Pitrelli, Beckman, & Hirschberg, 1994). In such exercises, several labelers, not necessarily from the same institution, analyze the same section of data, then the consistency of their labeling is determined statistically. Such experiments are very labor-intensive, but they show that the labels have some reality, in that they can be independently derived from acoustic data. However, the labels are generally not as reproducible as one would like, given that they are used as the basis of most of intonational phonology. Also, mere reproducibility does not necessarily establish that the labels have any linguistic reality, so questions persist.

We conclude that intonational phonologists and speech technologists would benefit from a technique that can connect impressionistic insights from intonational phonology with quantitative, objective data. Phonologists need methods that allow for empirical validation of labeling systems, and, if possible, access to larger bodies of data. Speech technologists require empirically tested and directly implementable data filtered by linguistic insights.

A first step in this direction was taken by Andruski and Costello (2004)[1] who used coefficients from polynomial equations, using Microsoft Excel, to explore small differences in the F0 contours of three low falling tones in Green Mong. Polynomial equations are a common mathematical approach to the description of curves. They are mathematical expressions involving a sum of powers in one or more variables multiplied by constants (e.g., $a_2x^2+a_1x+a_0$). They can conveniently produce a hierarchy of descriptions of increasing complexity and accuracy. 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 the Fujisaki model (Fujisaki, 1992), the TILT model (Taylor, 2000), and MOMEL (Hirst, di Christo, & Espesser, 1993). Curve-fitting also lies at the heart of the original Pierrehumbert (1980) model of intonational phonology. Pierrehumbert modeled intonation as a series of points connected by interpolations. Interpolations are straight lines or curves and can be generated by polynomial equations.

The language investigated by Andruski and Costello (2004) was Green Mong, a language spoken in South-East Asia in the region surrounding the Southern Chinese border. Green Mong has seven tones, three of which are quite similar in shape. They are low falling but differ in phonation type. Andruski and Costello asked whether F0 contour shape alone could be used to identify the tones. They estimated linear and quadratic polynomial 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 shape of the curvature of the three tones. Subsequent analysis showed that the three tones could be discriminated 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. 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.

Intonational diversity in the British Isles is well established (Cruttenden, 1995, 2001; Grabe, 2002b; Jarman & Cruttenden, 1976; Knowles, 1978; Local, Kelly, & Wells, 1986; Mayo, Aylett, & Ladd, 1996; Pellowe & Jones, 1978; Rahilly, 1991; Sebba, 1993; Sutcliffe & Figueroa, 1992; Tench, 1990; Vizcaino-Ortega, 2000; Walters, 1999; Wells, 1982; Wells & Peppé, 1996). In an autosegmental-metrical analysis of intonation in the IViE corpus (Grabe, Post, & Nolan, 2001), a publicly available corpus of speech data,[2] labelers identified seven nuclear accent types[3] (Grabe, 2004; Grabe & Post, 2002). These are the accent types we have investigated in the present study.

---

[1] Grabe, Kochanski, and Coleman (2003) also used polynomial equations to describe F0 in complete intonation phrases, to investigate and compare global properties of statements and questions.

[2] <http://www.phon.ox.ac.uk/IViE/>. IViE = Intonational Variation in English.

[3] In the present study, we excluded H*,%, an accent represented with only three tokens. The design of our statistical analysis included four dependent variables. As a rule of thumb, we assumed that at a minimum, every cell must have more cases than there are dependent variables.

The IViE corpus was recorded between 1997 and 2002 and contains 36 hours of speech from seven urban varieties of English. Recordings were made 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. In total, 108 speakers took part. They were 16 years of age and the recordings were made in their secondary schools. The speakers had grown up near the school, and as far as possible, we recorded speakers whose parents were also local. All speakers took part in the same battery of tasks. They read a list of sentences, read a familiar fairy tale, they retold the fairy tale in their own words, played an interactive game (a map task) and took part in a five minute period of free conversation. Six male and six female speakers were recorded in each location. Subsequently, approximately five hours of speech data were annotated with autosegmental-metrical intonation transcriptions. More information on the corpus and the transcriptions is available in Grabe (2004). The autosegmental-metrical analyses were also reported in Grabe (2004).

In the present study, our question was the following: if we build mathematical models of F0 for each of the hand-labeled nuclear accent types, will they be statistically different, and in what ways?

2 Method

2.1 Autosegmental-metrical annotations

Our research was based on autosegmental-metrical analyses of 714 read sentences in the IViE corpus. The sentences were produced by three male and three female speakers from each of seven dialects. The sentences are listed in Appendix A.

The data consisted of fully voiced declaratives, *wh*-questions, polar questions, and declarative questions, read in isolation. Different utterance types were included to elicit a wide range of intonation contours from each dialect. The intonation patterns were labeled via a combination of auditory analysis and visual inspection of fundamental frequency traces, a standard approach in the field (e.g., Beckman & Ayers Elam, 1997; Ladd, 1996). Transcriptions were made using the xlabel tool, part of the ESPS/xwaves+ package developed by Entropic Research Laboratories, Inc. A complete transcription consisted of an audio file, a time-aligned fundamental frequency trace and time-aligned text files containing transcriptions of intonation patterns. Intonation patterns were transcribed using the IViE system,[4] an autosegmental-metrical intonation transcription system developed for multidialect transcription of intonational variation in English (Grabe, 2002a; Grabe, 2004; Grabe, Nolan, & Farrar, 1998).

[4] The IViE labeling guide is available at <http://www.phon.ox.ac.uk/IViE/guide.html>. Unlike ToBI accents (Beckman & Ayers Elam, 1997), all IViE accents are left-headed, there are no phrase accents and intonation phrase boundaries can be given one of three specifications: H% if pitch rises above the level specified by the final starred tone, % if pitch continues level, and L% if pitch moves downwards.

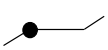
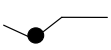
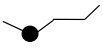
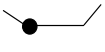
Accent labels were placed in the middle of the stressed syllable, by eye, using speech pressure waves and spectrograms following the IViE conventions, which are based on the ToBI conventions (Beckman & Ayers Elam, 1997). We determined subsequently that the accent marks were placed with a precision of 38 ms relative to syllable centers (5%-trimmed SD of a sample of 42 accents).

In the data, seven nuclear accent types were labeled in more than five instances. These accent types were included in the present study. They are listed, illustrated, and described in Table 1. As in Grabe (2004) and Grabe, Kochanski, and Coleman (2005), we do not separate downstepped accents from their nondownstepped counterparts.

The first column of Table 1 shows the labels, the second a stylized representation of a typical F0 shape associated with the label and the third a description using terminology commonly used in the British tradition of intonation analysis (e.g., Cruttenden, 1997). In autosegmental-metrical transcriptions, a nuclear accent consists of an utterance-final boundary tone, preceded by a tone accent (H*L, H*, L*H, or L*). To minimize confusion between accents and boundary tones, we have separated accent transcriptions (e.g., H*L) and boundary symbols (e.g., H%) by commas.

Table 1

Autosegmental metrical nuclear accent labels, stylizations, and descriptions of seven nuclear accents commonly observed in dialects of English spoken in the British Isles

Nuclear accent label	*Stylization*	*Description following the British tradition*
H*L,%[5]		Fall
H*L,H%		Fall-rise
H*,H%		High rise
L*H,%		Rise-plateau
L*H,H%		Rise
L*,H%		Late rise
L*H,L%		Rise-plateau-fall[6]

[5] The '%' boundary specification in IViE is based on Grabe (1998a) and is used in Grabe (1998b) and Grabe, Post, Nolan, and Farrar (2000). A % boundary indicates that the F0 of the last tone of the last accent in the intonation phrase is continued up to the boundary. The tonal specification of the last tone may be starred or unstarred. (NB. In Grabe, 1998a,b, % boundaries are given as '0%').

[6] Note that in nuclear accents comprising of only one or two syllables, the plateau may not be observed.

Table 2 shows the frequency of occurrence of each of the seven nuclear accents. The table illustrates the nature of nuclear accent variation in the IViE data. Some accent types were observed only in only one or two dialects (L*,H%, H*,H% and L*H,L%), some were observed in several dialects (e.g., L*H,%) and one was observed in all dialects (H*L,%). In addition, the frequency of accent types observed in more than one dialect varied. Although H*L,%, for instance, was observed in all dialects, the accent was rare in Belfast (observed in 2% of the cases) but common in Dublin (observed in 66% of the cases), this one instance demonstrating the great dialect variation even within "Irish English."

Table 2

Frequency of occurrence of each accent in percent, for each dialect

	London	*Cambridge*	*Bradford*	*Leeds*	*Newcastle*	*Belfast*	*Dublin*
H*L,%	46%	53%	57%	54%	50%	2%	66%
H*L,H%	18%	13%	3%	3%	4%	0%	5%
H*,H%	13%	0%	0%	1%	1%	0%	0%
L*H,%	0%	0%	34%	33%	43%	89%	22%
L*H,H%	10%	35%	1%	8%	0%	6%	1%
L*H,L%	0%	0%	0%	0%	0%	3%	6%
L*,H%	14%	0%	0%	0%	0%	0%	0%

Table 3 shows the total number of tokens of each nuclear accent types in the data set used in the present study.

Table 3

Distribution of nuclear accents in the sentence data in the IViE corpus

Nuclear accents	*Tokens*
H*L,%	414
L*H,%	187
H*L,H%	41
L*H,H%	32
H*,H%	15
L*,H%	12
L*H,L%	9
	710

Table 3 shows that the frequency distribution of nuclear accent types was uneven. In large speech corpora, however, an uneven distribution of linguistic types is the norm, not the exception. This property, where some features occur frequently but the vast majority are rare, was first noted in language by Zipf (1932); van Santen (1994) called it *lopsided sparsity*. The type count of rare features, however, can be so large that the likelihood of encountering one in a small sample is high: there will often be that one odd intonation pattern in each corpus. Consequently, we cannot dismiss rarer patterns as unimportant. 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 modeling

We will now give a brief description of our approach to polynomial modeling. Andruski and Costello (2004) found that three coefficients, describing the starting point of the contour, the slope and the curvature, were sufficient to account for the differences between three low falling tones in Green Mong. In our investigation of nuclear accent shapes in dialects of English, we raised the number of coefficients to four. The extra coefficient was required to model complex accents such as the Southern British English rise, which can take the shape of a staircase (cf. L*H,H% in Table 1). Data from our earlier investigations of the corpus also showed that we were likely to need the average. The labels suggested the presence of several rising accents. Some of these appeared to be distinguished solely by the location of the rise in the speakers' F0 range. Including the average allowed us to model accent differences of this type.

A further difference between our approach and Andruski and Costello's involved speaker normalization. Andruski and Costello's study was based on 270 sentences from six speakers. Our data consisted of 710 sentences produced by 42 speakers (6 speakers from each of 7 dialects). We normalized the data by each speaker's mean F0 (details are given below). Another difference is that we weighted the data so that the polynomial fits would emphasize loud, sonorant regions of the speech (e.g., syllable centers). This preferentially uses data where F0 measurements are more reliable, and where F0 may be more perceptually important.

Further, we fitted the data with *orthogonal* polynomials so that there would be minimal correlations among the coefficients that describe utterances. If one fits uniformly sampled data with polynomials of the form $a_2x^2+a_1x+a_0$, one will typically find that the resulting a_0, a_1, a_2 coefficients are far from independent of each other. This comes about because (in this example) both a_0 and a_2 raise the F0 value predicted by the model, so (for example) a given value of F0 might be explained by either a large a_0, or a large a_2, but not both. These correlations would otherwise need to be taken into account in the statistical analysis.

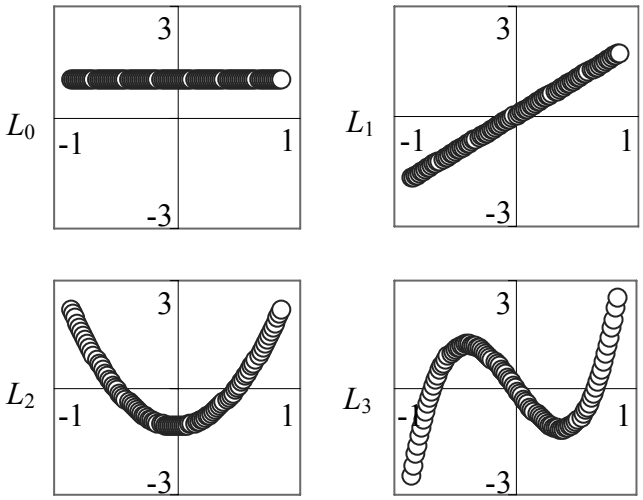
The orthogonal polynomials that we use are Legendre polynomials. We chose them (vs. other orthogonal polynomials) because they make the coefficients uniformly sensitive throughout the utterance. If one imagines making a change to the F0 data at different places in the utterance, then with Legendre polynomials the magnitude of the change in the coefficients will be the same, no matter whether the change was

made at the beginning, middle, or end. Other varieties, such as Chebyshev polynomials will emphasize certain regions of the utterance at the expense of others. The first four Legendre polynomials are shown in Figure 1.

Finally, unlike Andruski and Costello, we did not use Microsoft Excel. Our analysis was carried out with a set of custom-written Python scripts (Kochanski, Grabe, Coleman, & Rosner, 2005). A slightly simplified version of this approach could be carried out in Microsoft Excel, following the instructions given in the Appendix of Andruski and Costello (2004) following the methods presented in Appendix C.

Figure 1

Legendre polynomials L_0 to L_3. L_0 models the average of a given stretch of data, L_1 models the slope of the data, L_2 models the data as a parabola and L_3 models the data as a wave-shape. If the sign is inverted, the patterns are flipped about the horizontal axis



Before the data were analyzed, they were inspected for gross errors in the F0 tracks. An automated procedure was run to identify likely problem areas, and then a human labeler (one of the authors) inspected those (and other) areas and sometimes marked a change. The typical adjustment covered a small region, on average 56 milliseconds long, and the majority of the changes consisted of marking a region as unvoiced or irregularly voiced. Of the remainder, most were octave shifts. In all, less than 1% of the F0 or voicing data were adjusted, though 20% of the utterances were corrected. Then we analyzed fundamental frequency, using the loudness and periodicity signals to weight the importance and reliability of different regions. Details of the technique are given in Kochanski, Grabe, Coleman, and Rosner (2005).

We analyzed only the voiced region of the nuclear accent. In the section of the IViE corpus investigated here, all utterances were designed to be fully voiced. However, 12 utterances per dialect (*They are on the railings*) ended in a voiced fricative. In those utterances, the voiced fricative often became voiceless near the end. A small number of other utterances also ended with short irregularly voiced or voiceless regions.

In the corpus, each sentence consists of a single intonational phrase, so the analyzed region of F0 begins 100 milliseconds before the nuclear accent of the sentence (as defined by the final accent label preceding a boundary), and extends to the end of the voiced part of the sentence.

The central step in the data analysis was to represent the data as a best-fit sum of Legendre polynomials where each polynomial is normalized to have unit variance. The result of the analysis was a model for the F0 of each accent.

The model was specified by a set of coefficients, c_i, that multiply the different Legendre polynomials before they are added together:

$$M(x) = \sum_i c_i \cdot L_i(x) \tag{1}$$

This is similar to a Fourier analysis in that the low-ranking polynomials pick out slowly-varying properties and the higher-ranking polynomials pick out successively more rapidly varying properties (cf. Fig. 1).

Before we fit the polynomial model to the data, all F0 values were divided by the speaker's mean F0 then 1 was subtracted, so that a normalized F0 of zero corresponds to the speakers's mean F0. The time axis was also shifted and scaled so that the nuclear accent region corresponds to x values in Equation 1 between −1 and 1.

Next, a set of coefficients (suitable for use in Equation 1) was found that gave the best fit to the data. To find these, we used a weighted linear maximum-likelihood regression, exactly as in Kochanski et al. (2005). The first few coefficients have straightforward physical interpretations:

1. The first coefficient, c_0, is just the average F0 of the accent after normalization. So, if $c_0 = 0.1$, for example, the accent's average F0 is 10% higher than that speaker's average.
2. The second, c_1, is half the best-fitting slope of the accent, expressed as a fraction of the speaker's average F0 over the accent. So, $c_1 = -0.05$ corresponds to a modest (10%) decline in F0 over the accent. If $c_1 = 0$, there is no declining trend to an accent. (This doesn't rule out wiggles or even a sharp final fall if balanced by a suitable rise — such things appear in the higher coefficients.)
3. The third, c_2, corresponds to a broad dip or rise in the center of the accent. An accent with no overall curvature would have $c_2 = 0$.
4. Succeeding terms correspond to features of successively shorter duration. Figure 1 shows that the fourth coefficient, c_3 corresponds to a wave-like shape.

Then we carried out a multivariate analysis of variance (MANOVA) in SPSS, using the first four coefficients as dependent variables. LABEL was the independent variable.

3 Results

The statistical analysis showed that six of the seven hand-labeled nuclear accent types were statistically different. For transparency, we will refer to the dependent variables (i.e., the four coefficients) as AVERAGE (c_0), SLOPE (c_1), PARABOLA (c_2), and WAVE (c_3). LABEL (i.e., nuclear accent type) was the independent variable (H*L,%; H*L,H%; H*,H%; L*H,%; L*H,%; L*H,H%; L*H,L%). The analysis produced very highly significant main effects of LABEL on the dependent variables: AVERAGE, $F(1, 6) = 54.0$, $p < .001$; SLOPE, $F(1, 6) = 78.6$, $p < .001$; PARABOLA, $F(1, 6) = 14.4$, $p < .001$; WAVE, $F(1, 6) = 15.2$, $p < .001$.

Post hoc Tukey tests showed that every accent was significantly different from at least one other accent in one or more of the four coefficients. Two accent pairs differed at the 5% level (L*H,% vs. L*H,H% and H*L,% vs. L*H,L%), the rest were significant at $p < .01$. We did not find significant differences between L*,H% (the late rise observed in data from London), and the other two low rising accents L*H,% (the rise plateau) and L*H,H% (the rise). The full set of results is given in Tables 4 to 7.

Tables 4 to 7: Four half-matrices showing accent pairs distinguished significantly by the factors AVERAGE, SLOPE, PARABOLA, and WAVE. Black cells show significant differences at $p \leq .01$, gray cells show $p \leq .05$.

Table 4

AVERAGE	L*,H%	L*H,H%	L*H,%	L*H,L%	H*L,H%	H*L,%	H*,H%
H*,H%							
H*L,%							
H*L, H%							
L*H,L%							
L*H,%							
L*H,H%							
L*,H%							

Table 5

SLOPE	L*,H%	L*H,H%	L*H,%	L*H,L%	H*L,H%	H*L,%	H*,H%
H*,H%							
H*L,%							
H*L, H%							
L*H,L%							
L*H,%							
L*H,H%							
L*,H%							

Table 6

PARABOLA	L*,H%	L*H,H%	L*H,%	L*H,L%	H*L,H%	H*L,%	H*,H%
H*,H%							
H*L,%							
H*L, H%							
L*H,L%							
L*H,%							
L*H,H%							
L*,H%							

Table 7

WAVE	L*,H%	L*H,H%	L*H,%	L*H,L%	H*L,H%	H*L,%	H*,H%
H*,H%							
H*L,%							
H*L, H%							
L*H,L%							
L*H,%							
L*H,H%							
L*,H%							

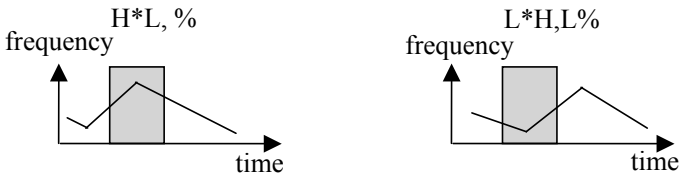
In the following paragraphs, we describe the results of the statistical analysis for each of the seven accents. For expository purposes, we assume that one significant difference between accents is sufficient for us to assume that accent patterns labeled as different have different acoustic manifestations in F0. (Note, however, that Tables 4–7 show that most accent pairs were distinguished by more than one statistically significant difference). In each case, we begin the descriptions of significant differences with the simplest: the lowest coefficient, the average.

The first accent listed in the leftmost column of each half-matrix is H*,H%. The data showed that H*,H% had an exceptionally high average F0 (cf. Fig. 3a, below). The statistical effect is shown in Table 4. The average was sufficient to distinguish the H*,H% accent from all other accents.

The accent below H*,H* in the leftmost column of Table 4 is H*L,%. H*L,% has a particularly low average. The low average distinguished H*L,% from the other accents with one exception: the rise-plateau fall L*H,L%, which was observed only in the data from Belfast and Dublin. Tables 5 and 6 do not show significant differences between H*L,% and L*H,L% either. The accents were not distinguished by SLOPE or PARABOLA. Table 7 shows that H*L,% and L*H,L% were distinguished by the fourth coefficient, WAVE, at the 5% level. This result is not surprising; H*L,% and L*H,L% have very similar contour shapes, and the difference between these accents lies solely in their alignment with the stressed syllable. A stylized illustration of the alignment difference is given in Figure 2. When the postnuclear stretch (i.e., the stretch following but not including the accented syllable) consists of no more than two syllables, the F0 contours exhibited by these accents can look very similar.

Figure 2

Stylizations of H*L,% and L*H,L%. The gray box indicates the location of the stressed syllable



The average also distinguished some of the remaining accents. The fall-rise H*L,H% had a significantly lower average than the plain rise L*H,H% and the rise-plateau L*H,% (see Fig. 3c, e, and f for an illustration). Table 5 shows that H*L,H% was distinguished from the late rise L*,H% by the SLOPE. Table 6 shows that H*L,H% was distinguished from the rise-plateau-fall L*H, L% by the third coefficient, PARABOLA.

In Tables 4 to 7, results for the rise-plateau-fall L*H,L% are shown below those for the fall-rise H*L,H%. L*H,L% was distinguished from L*H,% and L*H,H% by the AVERAGE. The SLOPE distinguished the rise-plateau-fall from the late rise L*,H%. L*H,H%, the plain rise, was distinguished from the rise-plateau L*H,% by the third coefficient, PARABOLA ($p \leq .05$). We did not find significant differences between the plain rise L*H,H% and the late rise L*,H%, nor between the rise-plateau L*H,% and the late rise L*,H%. We comment on this finding in the discussion.

Figure 3

Four-coefficient F0 profiles for the seven main nuclear accents in the IViE corpus. The coefficients are listed on the *x*-axis. The *y*-axis shows units of normalized F0 (0.1 = 10% of the speaker's average F0)

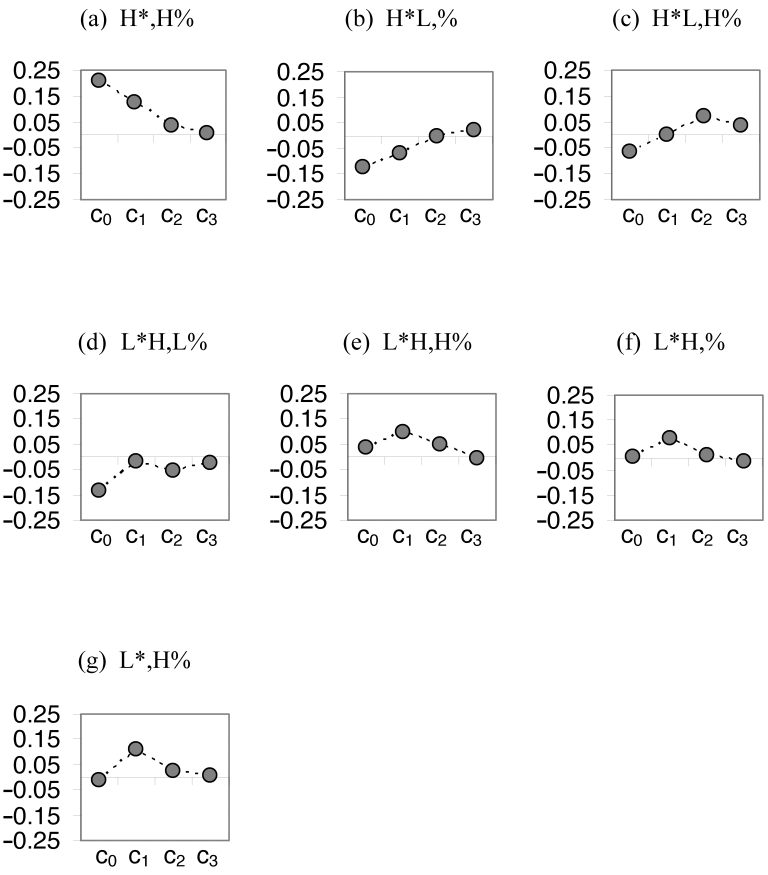


Figure 3 shows mean coefficient values for each nuclear accent type. In this figure, a negative first coefficient is equivalent to a low average F0 for the accent type; a positive first coefficient shows the opposite. A negative second coefficient indicates a falling slope, a positive second coefficient a rising slope. A negative third coefficient models a cup-shape, a positive third coefficient describes a domed shape. A negative fourth coefficient describes a falling-rising-falling component of the shape and a positive fourth coefficient a rising-falling-rising component. We will describe a few selected examples.

Figure 3a shows the four-coefficient representation of H*,H%. The first coefficient is positive and large: H*,H% accents had a relatively high average. The second coefficient is also positive and large: H*,H% accents have rising slopes. The fourth coefficient is close to zero, indicating that the WAVE component contributes very little to the shape.

Figure 3d shows the coefficients for rise-plateau-falls (L*H,L%). Averages are low and the slope contributes little to the shape of these accents. The fourth coefficient is not large, but recall that Figure 3 shows a marginally significant difference between the fourth coefficients of H*L,% and L*H,L%.

Figure 3e shows the plain rise L*H,H%. The second coefficient is large, as expected for a rising slope. The positive third coefficient shows the average accent of this class is somewhat cup shaped, rather than being strictly linear. Rises in Southern British English can take a staircase shape involving a rise on or following the stressed syllable, a short plateau and a sharp final rise at the end of the utterance (Cruttenden, 1997 or Ladd, 1996). The average of the fourth coefficient is about zero. We included the fourth coefficient to model staircase patterns in L*H,H% but Figure 3e shows that in our data, the fourth coefficient contributed little to the shape of L*H,H% (but recall that Fig. 3 only shows average coefficient values).

Figure 3g shows the late rise L*,H%. This accent was not distinguished significantly from the rise L*H,H% in Figure 3e nor from the rise-plateau L*H,% (Fig. 3f). We comment on this finding in the discussion.

Table 8

How distinct are the different accents? The table shows how many statistically highly significant distinctions can be made between each accent and others. Twenty-four comparisons are made for each accent (4 coefficients times 6 other accents). The rightmost column counts how many times each accent appears in the corpus

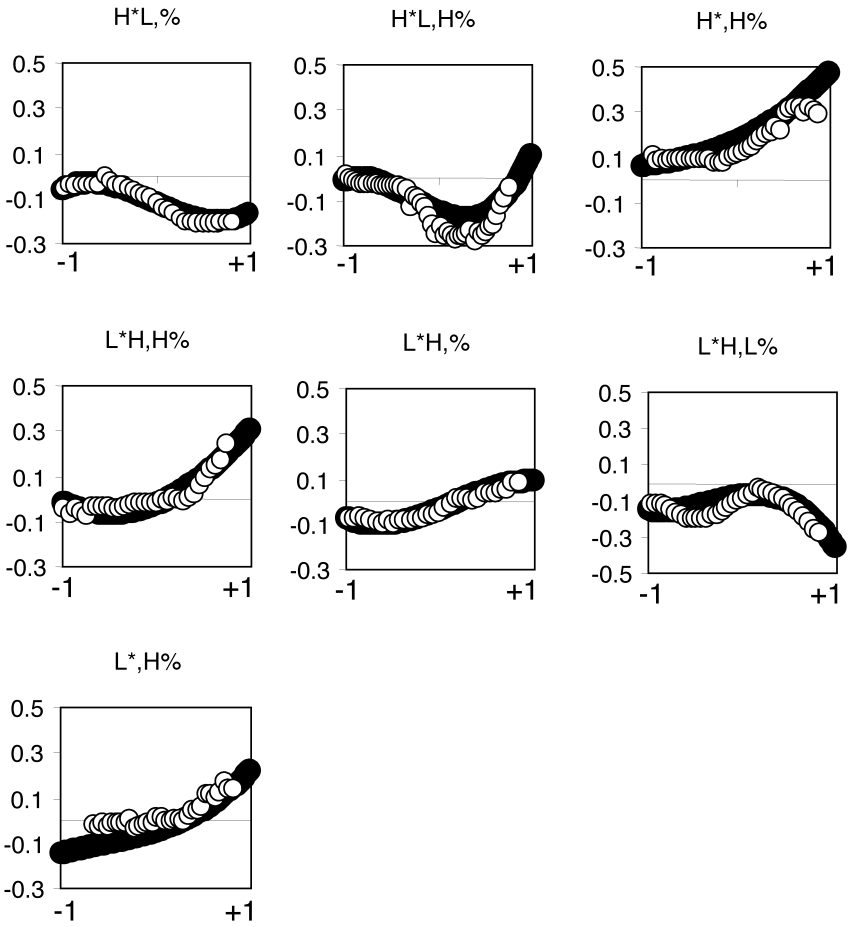
Nuclear accents	*Highly significant distinctions ($p<.01$)*	*Tokens*
H*L, %	14	414
H*L, H%	13	41
L*H, H%	11	32
L*H, L%	11	9
L*H, %	9	187
H*, H%	9	15
L*, H%	5	12

Table 8 takes us back to the notion of lopsided sparsity; it shows the number of highly significant distinctions between the coefficients that characterize each accent in the set and supports the idea that accents with a lower frequency of occurrence cannot be safely ignored. For example, there are only nine tokens of L*H,L%. Nevertheless, we found 11 statistically highly significant distinctions between this accent and the other accents in the set. H*L,%, on the other hand, is the most frequent accent in the data set and for H*L,%, we observed 14 highly significant distinctions.

To illustrate the plausibility of the orthogonal polynomial descriptions, we show an F0 model for each accent shape, reconstructed from the four coefficients in Figure 4.

Figure 4

Reconstruction of F0 models from the four coefficients (thick black lines) with superimposed F0 data (unfilled circles). The *x*-axis shows normalized time (−1 = beginning of accent; +1 = end of accent). The *y*-axis shows normalized F0



In Figure 4, the reconstructed F0 models (thick black lines) summarize the salient characteristics of each accent type. The reconstruction is 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, normalized F0 trace from the IViE corpus (unfilled circles) in each panel. This superimposed trace is the data that has the least mean-square difference to the model, and it shows that the models — despite being an average — are representative of the data.

In addition to being statistically significant, some of the differences are quite large and should be perceptually obvious. For instance, if we compare the H*L,% and H*L,H% accents (which are not the most different pair), we can see that at the end of the utterance, they differ by 0.2 normalized F0 units. For a speaker with a 170 Hz mean F0, this would be a difference of 34 Hz, substantially larger than segmental perturbations and the psychophysical just-noticeable-difference.

The models in Figure 4 provide a visual yet 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 (and with the exception of L*,H%, all contours are statistically different).

4 Discussion and Conclusion

We translated autosegmental-metrical intonation labels into four-term mathematical models of F0. Then we asked whether these would provide empirical validation of hand-labeled nuclear accents. A statistical analysis showed that six of seven accent types detected by hand-labelers were significantly different. This finding shows that although hand-labels may be impressionistic, they are associated with different F0 patterns.

The results of our statistical analysis also show that a model based on three coefficients can distinguish between the nuclear accent contours in our corpus. We found significant differences between the contours in the fourth coefficient, but in this data, at least, the information was redundant. The contribution of the fourth coefficient may, however, matter in other languages or in dialects not included in the present study. Further research is required.

One accent of the seven was not statistically different from two others; the late rise L*,H% could not be distinguished from L*H,H% or from L*H,%. This result requires consideration. We could conclude that the late rise L*,H% is not a separate accent type. In that case, L*,H% could be collapsed with another type of rise. The results of the statistical analysis do not indicate, however, which accent the late rise should be collapsed with. Moreover, we worked with very few tokens of L*,H%. Consequently, we cannot dismiss the issue of data sparsity. We worked with 12 tokens of L*,H%, versus 32 tokens of L*H,H% and 187 tokens of L*H,%. Since the analysis looks for statistical differences between the means of the distributions of coefficients associated with different labels, the analysis becomes more sensitive for labels with more data, because the means become more precisely defined as more measurements are made. One could argue that, had we worked with a larger number of L*,H% accents, a significant difference might have emerged. However, this argument points out a

limitation of the purely statistical analysis, as it can be shown that one can imagine any difference between the coefficients of two labels, however small, and make it statistically significant if one had a large enough corpus. For example, given a corpus 100 times larger than ours (an expensive project, but one that is certainly possible), F0 differences between labels that were smaller than 1 Hz, and thus imperceptibly small, could easily turn out to be statistically significant. One then might have a statistically significant difference between two labels that was too small to hear: such a difference could never be used to form a minimal pair. Thus, statistical significance is only meaningful if it is coupled with an estimate of the size of the effect (see our discussion of Fig. 4). Again, further research is required.

The rise-plateau L*H,% (commonly observed in Northern Irish English) and the rise L*H,H% only differed at the 5% level, in the third coefficient (PARABOLA). However, this result does not lead us to question the existence of two distinct rising patterns. First, the existence of two distinct patterns is well documented in the literature (e.g., Cruttenden, 1995, 1997; Lowry, 2001). Second, two distinct patterns have been observed in the IViE corpus, on texts in which the nuclear syllable is followed by two or more syllables (Grabe, 2004). An example is shown in Figure 5.[7] Rather, this finding shows that we need to consider the effect of neutralization on F0 modeling of nuclear accents (NB. The following comments also apply to the late rise L*,H%, discussed above). Distinctions between L*H,H%, and L*H,% can be observed clearly only if the accented syllable is followed by more than one syllable. On one or two syllables, the F0 patterns associated with L*H,% and L*H,H% are probably largely the same. In our materials (read sentences), the nuclear accent was usually placed on the last content word of the utterances (e.g., in the test sentence *We arrived in a Limo*, the nuclear accent was placed on *Li-* in *Limo*). Consequently, the accent rarely covered more than two syllables. Grabe, Post, Nolan, and Farrar (2000, p.167) show F0 patterns for rises on two-syllable words produced by speakers from Cambridge and Belfast. The contours do not differ substantially. Grabe et al. also show that in Belfast English, F0 patterns are truncated from the right. On one-syllable words, only the rise remains and there is no evidence of a plateau.

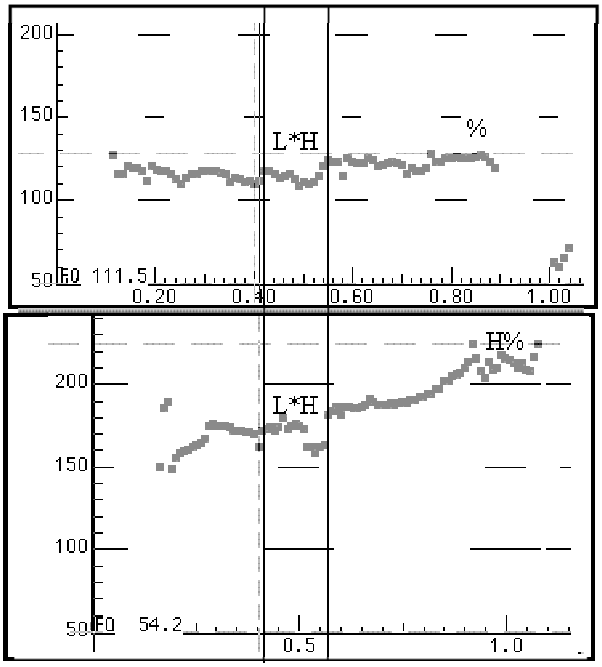
Finally, although we have shown that most accents are distinct, in the sense that the mean acoustic properties differ, and noted that some of the differences in F0 are large enough to be easily perceptible, we have not addressed the variability in the F0 patterns that characterize each accent. It is possible that the mean properties of all accents in one class (e.g., H*L,%) might be significantly and substantially different from accents in another class (e.g., L*H,L%), but that the two classes might be so variable that they overlap. One might then find a member of each class sharing the same acoustic properties. In such a case, a listener might not be able to decide which class the accent belonged to. This is a different and more stringent question from the means of the distribution; not only do we ask whether the means are different but also whether instance-to-instance variability is small (for the sake of this argument, we

[7] Note that in Northern Irish English, rises, and rise-plateaux co-occur. The co-occurrence can be taken as evidence for a phonological distinction between L*H,% and L*H,H% (Fletcher, Grabe, & Warren, 2002).

Figure 5

Realizations of L*,H% (top panel), produced by a male speaker from Belfast and L*H,H% (bottom panel), produced by a male speaker from Cambridge. The sentence was *May I leave the meal early?* In both utterances, the nuclear accent was placed on *leave*. The vertical lines in the picture show the location of the accented syllable

Nuclear L*H,% and L*H,H% produced over five syllables



do not consider the question of uneven distributions of accents in different dialects and the effect of uneven distributions on listeners' expectations).

We tested this question by building automated classifiers that operated from the orthogonal polynomial coefficients and measuring their performance, using the approach taken in Kochanski, Grabe, Coleman, and Rosner (2004). Details are given in Appendix D. We built and tested classifiers for each pair of accents. If they performed well, it would be proof that a certain pair of accents were acoustically distinct; if the classifier performed poorly, it would provide evidence that F0 alone was insufficient to distinguish these accent classes. The results showed that seven of the 21 accent pairs were sufficiently distinct that a listener could usefully separate them based on a single instance. Four pairs of accents that had distinct means were not separable; their acoustic properties overlap (as far as F0 is concerned) and our analysis procedure cannot distinguish those pairs based on a single instance. It is at least possible, however, that human listeners do not reliably make those distinctions based on F0 information alone; they may be using other acoustic information in addition to F0 or lexical context. The frequency of particular accent patterns in particular dialects and associated listener expectations may also have an effect on perception.

Our approach has a number of applications. First, as already discussed, the approach translates impressionistic descriptions of intonation into quantitative models of F0. Linguists can use these to investigate empirically the acoustic basis of phonological classifications. Second, the models are of value to speech technologists. Since the models are based on insights from linguistics, they are, in a sense, prefiltered. Hand-labelers 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. Third, at least potentially, the approach may provide linguists with access to larger bodies of data. In collaboration with speech technologists, intonation phonologists could develop methods that allow for automatic classification of large numbers of accents.[8] 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.

Our approach can also add to work on the alignment of intonation with segmental anchors, that is, vowels, consonants, and syllable boundaries (Arvaniti & Ladd, 1995; Arvaniti, Ladd, & Mennen, 1998, 2000; Atterer & Ladd, 2004; d'Imperio, 2001; Frota, 2002; Grabe, Post, Nolan, & Farrar, 2000; Ladd, Mennen, & Schepman, 2000; Prieto, van Santen, & Hirschberg, 1995; Silverman & Pierrehumbert, 1990). Polynomial models of F0 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. The experimenter needs to mark the section of speech under investigation (a vowel, a vowel-consonant combination, a syllable or a word) and place the section on a normalized time axis, between −1 and +1. (Note that the normalization takes care of the effect of changes in speaking rate). The shape of F0 in the section can then be modeled. A stylized example illustrating how polynomial modeling can contribute to work on alignment is given in Appendix C. The approach illustrated in Appendix C has a number of advantages. Firstly, the experimenter is not exclusively restricted to hand-measurements of peaks and valleys. Sometimes, peaks and valleys in F0 are not well defined and awkward decisions are required. Secondly, curve-fitting can bridge some voiceless regions. Our current technique handles 10% voiceless regions in the F0 signal cleanly. Consequently, experimenters do not have to restrict themselves to carefully designed, fully voiced laboratory speech. Finally, the approach provides information at more than one level. For instance, hand-measurements of peak alignment do not reveal consistent changes in the average of the contour; a peak may appear later in some accents but it may also be lower. Alternatively, the peak may appear later but the slope of the contour may not be affected. Polynomial models provide the data required to test such observations.

More generally, the approach allows for a combination of qualitative and quantitative comparisons of intonation systems across dialects and languages. Cross-

[8] Note that the location of stressed syllables would have to be hand-labeled or determined automatically using procedures such as the ones developed by, for example, Streefkerk, Pols, and ten Bosch (1998), Tamburini and Caini (2005) or Kochanski, Grabe, Coleman, and Rosner (2005).

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 will provide new insights.

We conclude that polynomial modeling 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 F0 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.

manuscript received: 07. 22. 2005
manuscript accepted: 06. 20. 2006

References

ABRAMOWITZ, M., & STEGUN, I. A. (1965). *Handbook of mathematical functions*. New York: Dover Publications, ninth printing (1970).

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.

ARVANITI, A., & LADD, D. R. (1995). Tonal alignment and the representation of accentual targets. *Proceedings of the 13th International Congress of Phonetic Sciences*, vol. **4** (pp. 220–223). Stockholm.

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.

ARVANITI, A., LADD, D. R., & MENNEN, A. (2000). What is a starred tone? Evidence from Greek. In M. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology V*: *Acquisition and the lexicon* (pp.119–131). Cambridge: Cambridge University Press.

ATTERER, M., & LADD, D. R. (2004). On the phonetics and phonology of "segmental anchoring" of F0: Evidence from German, *Journal of Phonetics*, **32**, 177–197.

BECKMAN, M. E., & AYERS ELAM, G. (1997). *Guidelines for ToBI Labelling, version 3*. The Ohio State University Research Foundation, Ohio State University.

CRUTTENDEN, A. (1995). Rises in English. In J. Windsor Lewis (Ed.), *Studies in General and English Phonetics. Essays in honour of Prof. J. D. O'Connor* (pp.155–173). London: Routledge.

CRUTTENDEN, A. (1997). *Intonation*. Cambridge: Cambridge University Press.

CRUTTENDEN, A. (2001). Mancunian intonation and intonational representation. *Phonetica*, **58**, 53–80.

CUTLER, A., DAHAN, D., & van DONSELAAR, W. A. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, **40**, 141–202.

D'IMPERIO, M. (2001). Tonal structure and pitch targets in Italian focus constituents. *Speech Communication*, **33**, 339–356.

FLETCHER, J., GRABE, E., & WARREN, P. (2004). Intonational variation in four dialects of English: The high rising tune. In Jun, S. (Ed.), *Prosodic typology*. *The phonology of intonation and phrasing* (pp. 390–409). Oxford: Oxford University Press.

FROTA, S. (2002). Tonal association and target alignment in European Portuguese nuclear falls. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7* (pp. 387–418). Berlin: Mouton de Gruyter.

FUJISAKI, H. (1992). Modelling the process of fundamental frequency contour generation. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisasaka (Eds.), *Speech perception, production and linguistic structure* (pp. 313–328). Amsterdam: IOS Press.

GILI FIVELA, B. (2002). Tonal alignment in two Pisa Italian peak accents. In B. Bel & I. Marlin (Eds.), *Proceedings of the Speech Prosody 2002 conference* (pp. 339–342). Aix-en-Provence: Laboratoire Parole et Langage.

GRABE, E. (1998a). Comparative intonational phonology: English and German. *MPI Series in Psycholinguistics 7*, Wageningen: Ponsen en Looien.

GRABE, E. (1998b). Pitch accent realisation in English and German. *Journal of Phonetics*, **26**, 129–144.

GRABE, E. (2002a). *The IViE labelling guide*. Internet document. <http://www.phon.ox.ac.uk/IViE/guide.html>.

GRABE, E. (2002b). Variation adds to prosodic typology. In B. Bel & I. Marlin (Eds.), *Proceedings of the Speech Prosody 2002 conference*. Aix-en-Provence, Laboratoire Parole et Langage (pp. 127–132). Aix-en-Provence.

GRABE, E. (2004). Intonational variation in English. In P. Gilles & J. Peters (Eds.), *Regional variation in intonation* (pp. 9–31). Tübingen: Niemeyer.

GRABE, E., KOCHANSKI, G., & COLEMAN, J. (2003). Quantitative modelling of intonational variation. *Proceedings of speech analysis and recognition in technology*, *linguistics and medicine 2003*. [NB. The publication process of the volume has been delayed. The paper is available at <http://www.phon.ox.ac.uk/oxigen/publications.php>].

GRABE, E., KOCHANSKI, G., & COLEMAN, J. (2005). The intonation of native accent varieties in the British Isles — potential for miscommunication? In K. Dziubalska-Kolaczyk & J. Przedlacka (Eds.), *English pronunciation models: A changing scene* (pp. 311–337). Linguistic Insights Series: Peter Lang.

GRABE, E., NOLAN, F., & FARRAR, K. (1998). IViE — A comparative transcription system for intonational variation in English. *Proceedings of the 5th International Conference on Spoken Language Processing* (ICSLP), 1998. Sydney: Australian Speech Science and Technology Association, Incorporated (ASSTA).

GRABE, E., & POST, B. (2002). Intonational variation in English. In B. Bel & I. Marlin (Eds.), *Proceedings of the Speech Prosody 2002 conference* (pp. 343–346). Aix-en-Provence: Laboratoire Parole et Langage.

GRABE, E., POST, B., & NOLAN, F. (2001). *The IViE corpus*. Electronic resource available from <http://www.phon.ox.ac.uk/IViE>.

GRABE, E., POST, B., NOLAN, F., & FARRAR, K. (2000). Pitch accent realisation in four varieties of British English. *Journal of Phonetics*, **28**, 161–185

GRICE, M., REYELT, M., BENZMÜLLER, R., MAYER, J., & BATLINER, A. (1996). Consistency in transcription and labelling of German intonation with GToBI. *Proceedings of the 4th International Conference on Spoken Language Processing* (ICSLP) (pp. 1716–1719). Philadelphia, PA.

HIRST, D. J., di CRISTO, A., & ESPESSER, R. (1993). Levels of representation and levels of analysis for the description of intonation systems. In M. Horne (Ed.), *Prosody: Theory and experiment* (pp. 51–88). Dordrecht: Kluwer Academic Publishers.

JARMAN, E., & CRUTTENDEN, A. (1976). Belfast intonation and the myth of the fall. *Journal of the International Phonetic Association*, **6**, 4–12.

JUN, S.-A., SOOK-HYANG, L., KEEHO, K., & YONG-JU, L. (2000). Labeler agreement in transcribing Korean intonation with K–ToBI. *Proceedings of the 6th International Conference on Spoken Language Processing* (ICSLP), Beijing, China.

KNOWLES, G. O. (1978). The nature of phonological variables in Scouse. In P. Trudgill (Ed.), *Sociolinguistic patterns in British English* (pp. 8–91). London: Edward Arnold.

KOCHANSKI, G., GRABE, E., COLEMAN, J., & ROSNER, B. (2005). Loudness predicts prominence; fundamental frequency lends little. *Journal of the Acoustical Society of America*, **118**, 1038–1054.

KOCHANSKI, K., & SHIH, C. (2003). Prosody modeling with soft templates. *Speech Communication*, **39**, 311–352.

LADD, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

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.

LOCAL, J., KELLY, J., & WELLS, W. (1986). Towards a phonology of conversation: Turn-taking in urban Tyneside speech. *Journal of Linguistics*, **22**, 411–437.

LOWRY, O. (2001). *Belfast intonation patterns*: *Testing the ToBI Framework of Intonational Analysis.* PhD thesis: University of Ulster.

MAYO, C., AYLETT, M., & LADD, D. R. (1997). Prosodic transcription of Glasgow English: An evaluation study of GlaToBI. In A. Botinis, G. Kouroupetroglou & G. Carayiannis (Eds.), *Proceedings of an ESCA Workshop*. *Intonation*: *Theory, Models and Applications* (pp. 231–234). European Speech Communication Association.

OHALA, J. (1983). Cross-language use of pitch: An ethological view. *Phonetica*, **40**, 1–18.

PELLOWE, J., & JONES, V. (1978). On intonational variability in Tyneside speech. In P. Trudgill (Ed.), *Sociolinguistic patterns in British English* (pp.111–113). London: Arnold.

PIERREHUMBERT, J. B. (1980). *The phonology and phonetics of English intonation*. Doctoral dissertation, Cambridge, MA: MIT.

PIERREHUMBERT, J., & HIRSCHBERG, J. (1990). The meaning of intonation contours in the interpretation of discourse. In P. R. Cohen, J. Morgan & M. E. Pollack (Eds.), *Intentions in communication* (pp. 271–311). Cambridge, MA: MIT Press

PITRELLI, J., BECKMAN, M., & HIRSCHBERG, J. (1994). Evaluation of prosodic transcription labelling reliability in the ToBI framework. *Proceedings of the Third International Conference on Spoken Language Processing* (ICSLP), vol. **2**, (pp. 123–126). Yokohama.

PRIETO, P., van SANTEN, J., & HIRSCHBERG. J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics*, **23**, 429–451.

RAHILLY, J. (1991). *Intonation patterns in normal hearing and postlingually deafened adults in Belfast*. PhD thesis, The Queen's University, Belfast.

SEBBA, M. (1993). *London Jamaican*: *Language systems in interaction*. London: Longman.

SILVERMAN, K., & PIERREHUMBERT, J. (1990). The timing of prenuclear high accents in English. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 72–106). Cambrige: Cambridge University Press.

STREEFKERK, B. M., POLS, L. C. W., & ten BOSCH, L. F. M. (1998). Automatic detection of prominence (as defined by listeners' judgments) in read aloud Dutch sentences. *Proceedings of the 5th International Conference on Spoken Language Processing* (ICSLP), vol. **3**, (pp. 683–686). Sydney.

SUTCLIFFE, D., & FIGUEROA, J. (1992). *System in black language*. Clevedon, Avon: Multilingual Matters.

TAMBURINI, F., & CAINI, C. (2005). An automatic system for detecting prosodic prominence in American English continuous speech. *International Journal of Speech Technology*, **8**, 33–44.

TAYLOR, P. (2000). Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, **107**, 1697–1714.

TENCH, P. (1990). The pronunciation of English in Abercrave. In N. Coupland (Ed.), *English in Wales* (pp. 13–141). Clevedon, Avon: Multilingual Matters.

Van de WEIJER, J. (1997). Language input to a prelingual infant. In A. Sorace, C. Heycock & R. Shillcock (Eds.), *Proceedings of the GALA 1997 Conference on Language Acquisition* (pp. 290–293). Edinburgh: Scotland.

Van de WEIJER, J. (1998). Language input and word discovery. PhD thesis, *MPI Series in Psycholinguistics 9*. Wageningen: Ponsen en Looien.

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.

VIZCAINO-ORTEGA, F. (2002). A preliminary analysis of yes/no questions in Glasgow English. In B. Bel & I. Marlin (Eds.), *Proceedings of the Speech Prosody 2002 conference* (pp. 683–686). Aix-en-Provence: Laboratoire Parole et Langage.

WALTERS, J. R. (1999). *A study of the segmental and suprasegmental phonology of Rhondda Valleys English*. PhD thesis, University of Glamorgan.

WEISBERG, S. (1985). *Applied linear regression*, 2nd edition. New York: John Wiley and Sons.

WELLS, B., & PEPPÉ, S. (1996). Ending up in Ulster: Prosody and turn-taking in English dialects. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation*: *Interactional studies* (pp. 101–130). Cambridge: Cambridge University Press

WELLS, J. C. (1982). *Accents of English*: Vol. **2**, *The British Isles*. Cambridge: Cambridge University Press.

ZIPF, G. K. (1932). *Selective studies and the principle of relative frequency*. Cambridge, MA: Harvard University Press.

Appendix A

Materials

(1) Simple statements: 1. We live in Ealing. 2. You remembered the lilies. 3. We arrived in a limo. 4. They are on the railings. 5. We were in yellow. 6. He is on the lilo. 7. You are feeling mellow. 8. We were lying.

(2) Questions without morphosyntactic markers: 1. He is on the lilo? 2. You remembered the lilies? 3. You live in Ealing?

(3) Inversion questions: 1. May I lean on the railings? 2. May I leave the meal early? 3. Will you live in Ealing?

(4) *wh*-questions: 1. Where is the manual? 2. When will you be in Ealing? 3. Why are we in a limo?

Appendix B

The analysis in this paper was based on three measures that are derived from the acoustic data:

1. A measure of the F0,
2. A measure of the loudness,
3. A measure of the periodicity of the voicing.

We used the loudness and periodicity measures to weight the importance and reliability of different regions of the F0 signal. The details can be found in Kochanski et al. (2005), but we will present a qualitative outline here.

The fundamental frequency, F0, was obtained using better-than-standard software, (ESPS get_f0), but not all F0 measurements are equally valuable. For instance, measurements in quiet regions of speech may be drowned out by background noise, and therefore will sometimes have no impact on the listener at all. Such noise is common: Kochanski et al. (2005) estimated that 50% of the U.S. urban population has enough noise indoors that regions of speech more than 15 dB below the peaks will be masked by environmental noise. Thus, one should reduce the weight of quiet regions of the data to mimic real world perception of speech.

Speech is not equally periodic. If the larynx is not vibrating uniformly, there will not be a unique oscillation frequency, and one might expect that the perceived pitch would be ill-defined. Our weighting function combines these two observations, and puts the greatest emphasis on loud, periodic regions, which are normally the centers of syllables. Our fits to the data, and thus the coefficients, primarily reflect changes in these high-weight regions.

In this study, the fundamental frequency data were normalized, to allow for comparisons between data from different speakers. The normalization was carried out separately for each speaker. We divided each F0 data point by the average frequency of the speaker and subtracted 1. A value of '0' then corresponds to the average frequency. A value of one corresponds to one octave above the speaker's average fundamental frequency.

Orthogonal Polynomials

A family of orthogonal polynomials is a set of mathematical functions that can be used to describe a curve. There are an infinite number of families of orthogonal polynomials, but they all share some common properties:

i. They can be used, reversibly, to analyze a curve and to reconstruct it.

ii. They form a complete set of functions, so that if you use enough functions from the family, you can reconstruct any curve to any desired accuracy.

iii. They are orthogonal, which means that, in a specific sense, they do not correlate with each other. Each function of a family can be used to measure a different property of a curve, and (in the common case) the measurements turn out to be (nearly) statistically independent of each other with (very nearly) Gaussian distributions.

Beyond those common properties, one can choose a family of orthogonal functions that is tailored for the desired analysis. Some families are smooth and continuous; others are not. Some families are composed of functions that capture information across the whole utterance; others contain functions that are each localized in a different little region. When one is using orthogonal polynomials to represent data, the family one chooses depends on the questions one wants to answer.

The IViE intonational labels are localized in the sense that they are associated with a particular syllable and that they primarily describe the intonation over a domain that is a syllable or two wide. We chose a smooth and continuous family of orthogonal polynomials to represent F0 within that domain, as we expect that the intentionally controlled aspects of intonation should be, by and large, smooth

and continuous (Kochanski & Shih, 2003, see especially Section 1.2) because F0 is controlled by muscle tensions which are smooth functions of time. Additionally, we chose a family of polynomials having a uniform sensitivity to the data across the whole region. As a result, we chose the family known as Legendre polynomials (Abramowitz & Stegun, 1970). The family of Legendre polynomials is ordered in terms of increasing wiggliness: The first Legendre polynomial is a constant, the second is a linear slope, the third is a parabola, and in general, the n^{th} Legendre polynomial has $(n-1)/2$ peaks and $(n-1)/2$ troughs, if we count a high (low) point at an edge of the utterance as half a peak (trough).

Analysis with Legendre Polynomials

The coefficients of the polynomials were determined using a weighted linear maximum a posteriori (MAP) regression (a variant of a "multivariate linear regression" in the statistics literature). The result is similar to a Fourier analysis in that the low-ranking polynomials pick out slowly-varying properties and the higher-ranking polynomials pick out successively more rapidly varying properties. One can say that the n^{th} Legendre polynomial picks out variations in F0 which have a scale of $2/N$ of an intonational phrase.

Given a family of functions, which we can write as $f_i(x)$, one can analyze an F0 curve $f(t)$, in terms of that family by standard techniques of multiple linear regression. First, $f(t)$ is normalized to compensate for interspeaker differences:

$$F(t) = f(t) \Big/ \bar{f} \quad - 1 \qquad \text{[B.2]}$$

where $\bar{f}$ is the speaker's F0, averaged over all the single-speaker utterances in the IViE database. A normalized F0 of 0.1 corresponds to an F0 that is 10% above the speaker's average.

Second, the time axis is linearly stretched and shifted so that it covers the range: (−1, 1)

$$F(x) = F(2 \cdot (t - t_c) / D), \qquad \text{[B.3]}$$

where t_c is the center of the intonational phrase and D is the accent's duration.

Third, a set of equations is determined, one for each accent's F0 contour. These equations are a model for the F0 contour, written as a sum of the orthogonal functions, each multiplied by a constant:[9]

$$M(x) = \sum_{i=0}^{N} c_i \cdot L_i(x), \qquad \text{[B.4]}$$

where c_i is the (as yet unknown) coefficient that shows how much the i^{th} function contributes to the shape of the F0 curve. The sum is taken over however many functions in the family that are needed. Each possible combination of values for the c_i gives a different model, so we must select the best of these many possible models. Thus, one

[9] Note that this is the same as Equation 1, except that B.4 has ranges specified.

computes the total error for each model and chooses the model that minimizes the error. The error is

$$\chi^2 = \sum_X \left(M(x) - F(x)\right)^2 \cdot W(x) + \lambda \cdot R\,, \qquad \text{[B.5]}$$

where X is the set of places where we have F0 measurements, $W(x)$ controls how much weight we give to errors in different places, and R is the regularization term. The regularization term is designed to keep the analysis well-behaved for utterances with substantial unvoiced regions, and has little effect on the data used in this paper.

The total error indicates which model is the best representation for the observed F0. Bearing in mind that each combination of coefficients gives a different model, we are simply taking the set of coefficients that gives the smallest χ^2. Linear regression just provides an efficient way to search for the best values for the coefficients.

The result of the analysis is a set of coefficients, c_0, c_1, c_2, …. The coefficients allow the data to be reconstructed by way of Equation B.4: one adds together the various basis functions multiplied by the coefficients. If one coefficient is particularly large, the data and the model will tend to have the shape of the orthogonal polynomial that is multiplied by that large coefficient.

Appendix C

This appendix provides an example of how polynomial modeling can contribute to work on the alignment of accents with words, syllables or vowels. The example was constructed in Microsoft Excel, following the instructions in Andruski and Costello (2004, p.139), with an addition: we converted the Excel coefficients to Legendre coefficients. These are easier to interpret: the first coefficient captures the average, the second the slope and the third captures the curvature. In the equations produced by Excel, each term represents a combination of various properties of a curve. The values required for the conversion from Excel and an explanation of the conversion process are given below. We restrict ourselves to the first three Legendre coefficients, for simplicity.

Figure 6 shows three stylized F0 traces. Traces 1 and 2 are similar but in Trace 2, the peak is displaced to the right, representing a later alignment of the peak. In Trace 3, the peak is in the same location as in Trace 1, but the peak height is lower. The x-axis shows units of normalized time. The y-axis shows F0 values[10].

We modeled the traces in Figure 6 as follows. In Excel, we fitted trendlines to the traces, using second order polynomial equations. Detailed instructions are given in Andruski and Costello (2004, p.139). Note that in Excel, equations for trend lines start with the highest order term. The resulting equations were the following:

i. Trace 1 $y = -46.917\,x^2 + 0.0003\,x + 41.148$

[10] NB. We have not attempted to place the sample traces into an imaginary speaker's F0 range, this is why the traces start at 0 Hz. Moreover, the traces shown in Figure 6 are not normalized, but the procedures in this Appendix also apply to normalized data.

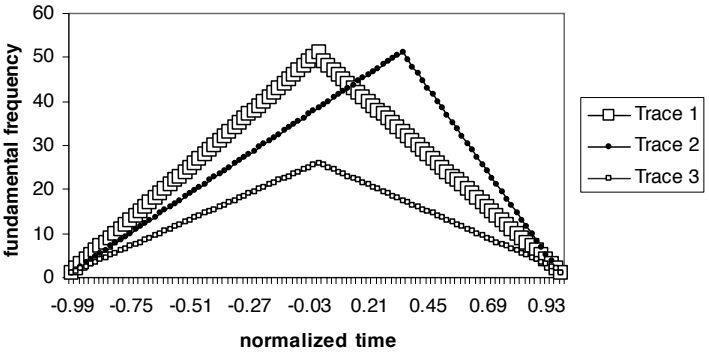


Figure 6

Sample F0 traces illustrating differences in alignment

ii. Trace 2 $y = -42.159\ x^2 + 8.437\ x + 39.797$

iii. Trace 3 $y = -23.476\ x^2 + 0.36\ x + 21.32$

The polynomial equations generated by Excel have the following shape:

iv. $y = a_0 + a_1\ x + a^2\ x^2$

Then we converted the coefficients given in Equations (i–iii) to Legendre coefficients. The goal was to find a sum of the first few Legendre polynomials that exactly matched equations i–iii. The Legendre coefficients appear as c_0, c_1, c_2 in Equation (v).

v. $y = c_0 L_0(x) + c_1 L_1(x) + c_2 L_2(x).$

The coefficients for the first three Legendre polynomials coefficients are defined as follows (cf. Fig. 1 above):

vi. $L_0(x) = 1$

vii. $L_1(x) = 1.732\ x$

viii. $L_2(x) = 3.355\ x^2 - 1.118$

Using equations (vi–viii), one can find a general rule for converting the coefficients obtained from Excel into Legendre coefficients:

ix. $c_0 = a_0 + 0.333\ a_2$

x. $c_1 = 0.577\ a_1$

xi. $c_2 = 0.298\ a_2$

(NB: equations (ix–xi) are generally valid and not restricted to this example. The only requirement for validity is that the time axis must range from −1 to 1. A similar set of four equations can be obtained if third-order polynomials are desired.)

As an example, we convert the Excel model for Trace 1 to Legendre coefficients:

Trace 1 $y = -46.917\ x^2 + 0.0003\ x + 41.148$

In this example $a_0 = 41.148$, $a_1 = 0.003$ and $a_2 = -41.917$

To calculate the first Legendre coefficient (the average) for Trace 1:

$c_0 = 41.148 + 0.333 \cdot (-46.917)$

and so on for c_1 and c_2.

Table 9 gives the results of the conversion.

Table 9

Coefficients	*Trace 1*	*Trace 2*	*Trace 3*
c_0	25.525	25.758	13.502
c_1	0.0	4.871	0.208
c_2	− 13.986	− 12.568	− 6.998

Figure 7 illustrates the results.

Figure 7

Modeling of phonetic detail in F0 with Legendre polynomials

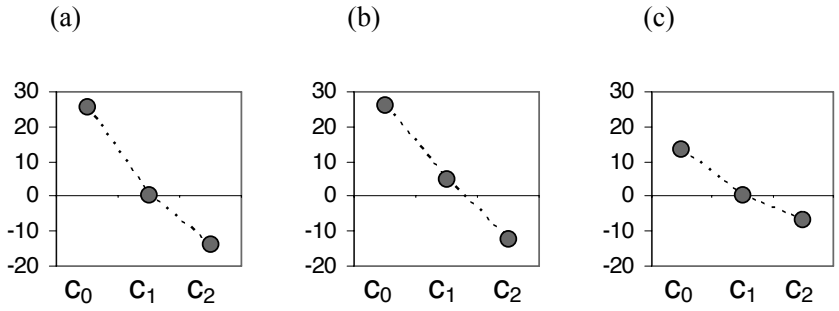


Figure 7(a) shows a Legendre model of Trace 1 in Figure 6. The first coefficient c_0 shows that the average F0 of the trace is 25 Hz. The second coefficient, c_1, shows that the overall slope is neither rising nor falling (since the trace is rising-falling and symmetrical). The third coefficient c_2 shows that the trace rises and then falls.

Figure 7(b) models Trace 2 in Figure 6. In this trace, the peak is displaced to the right. The first coefficient in Figure 7(b) shows that the average is largely unchanged. The second coefficient, however, has changed and is now positive. We learn from this that the trace is no longer symmetrical and that the rising slope is dominant over the falling slope (if the falling slope were dominant, the coefficient would be negative). The third coefficient, again, shows that the trace rises and then falls.

Figure 7(c) models Trace 3 in Figure 6. The alignment of the peak in Trace 3 is identical to that in Trace 1 but the average is lower. This is evident in Figure 7(c). The first coefficient shows that the average is lower, the second coefficient is again zero, as in Figure 7(a), indicating a centered, symmetric peak and the third shows that Trace 3 is domed but less so than Trace 1.

In sum, Appendix C shows how polynomial models can contribute to the investigation of fine phonetic detail in the alignment of F0. In our models, all stretches of F0 under investigation are placed on a normalized time scale. Consequently, as long as

one investigates differences in alignment on the same text, the models provide directly comparable information on the average, slope, and curvature of F0 patterns. Any change in peak alignment, for instance, will change slope and curvature. Parameters such as these are likely to be perceptually relevant. Andruski and Costello (2004), for instance, showed that the curvature of F0 was sufficient to distinguish between three falling tones in Green Mong. We also know that manipulations of average pitch are used for pragmatic purposes in speech (e.g., Ohala, 1983) and they play a role in Motherese (van de Weijer, 1997, 1998).

Appendix D

We also investigated instance-to-instance variability in the F0 patterns of the seven nuclear accents included in our study. In this appendix, we describe our approach. This approach differs from the MANOVA approach in the body of the paper which asks if the average of one group is different from the average of another group. Here, we investigate whether a single utterance provides enough information to reliably deduce its class (i.e., the IViE label).

For each pair of accents, we built automated classifiers that operated from the orthogonal polynomial coefficients and we measured the classifier performance. If they performed well, it would be proof that the two types of accents could be distinguished on the basis of a single example. This is a listener-oriented test: it asks whether a certain kind of listener could correctly deduce the intended phonological class of a particular accent. The MANOVA test can be thought of as being oriented more towards the speaker: it asks whether there is evidence that the speaker attempts to distinguish the two phonological classes.

We used the classifiers developed in Kochanski, Coleman, Grabe, and Rosner (2005). These generate a score, K, which ranges from zero to one; zero indicates that the classification is no better than chance and one indicates perfect classification based on the orthogonal polynomial coefficients. We tried both linear discriminant classifiers and more complex quadratic discriminant classifiers on data sets containing three and four orthogonal polynomial coefficients. All yielded essentially the same result; we report results for the linear discriminant classifier operating on the first four orthogonal polynomial coefficients.

The classifier has four main limitations relative to an ideal listener: (1) It is given less timing information than a human might be expected to have. All the classifier's data is anchored to the end of the utterance and the time of the nuclear accent label; recall that the label is not placed at any particular location within the accented syllable, so that any precise timing relationships that might be important are unfortunately lost. Further, an ideal listener might plausibly relate features in the F0 curve to other anchor points, such as the beginning of the final syllable or (where they exist) any syllable(s) between the accent and the final syllable; this data is not available to the classifier. (2) It does not consider possible prosodic cues like loudness, duration, and vowel quality. (3) It is not told the interval between the accent and the end of the sentence, either in terms of syllables or seconds, so that neutralization cannot be

modeled. (4) It does not have information on the segmental content of sentences, so segmentally-induced shifts in F0 cannot be compensated for.

Note that the "chance" classification depends upon the relative frequencies of the two classes. If the two classes have an equal number of instances, you expect to be able to guess the classification 50% of the time even without looking at the F0 data. However, if one of the classes is much more frequent than the other, one can guess the class more often than 50%. Take L*H,% versus H*,H% as an example: The corpus contains 187 instances of L*H,% but only 15 instances of H*,H%, so if one were trying to blindly guess the class of an accent that was known to be one or the other, one could be right 93% of the time by guessing L&H,%. So, "chance" here means how well you could do with knowledge of the accent frequencies, but without "hearing" an individual accent.

For each pair of accents, we randomly split the data 60–40 into a training set and a test set, produce five good classifiers for a given split, and repeat with 15 different splits. The 90 resulting classifiers each report a *K*-score; we summarize the results in Table 10. Note that *SD*s of *K*-scores were large since since some of our accents were represented by a small number of tokens. In Table 10, the number of tokens for each accent are given in brackets, following the accent symbol.

In Table 10, we mark those classifiers whose performance is significantly above zero (indicated by '>0') and those classifiers whose performance was significantly greater or less than $K=0.5$ (indicated by '>0.5' or '<0.5', respectively). We used a *t*-test with 15 degrees of freedom with the measured mean and SD for *K*. For the test against zero, we used a one sided *t*-test; for the test against 0.5, we used a two-sided *t*-test. All tests were at $p=.05$, and an asterisk marks $p<.01$.

The symbol '>0.5' means that the two types of accent being compared are well separated and that one can look at the four coefficients from a single instance and reliably deduce the accent's phonological class. Conversely, the symbol '<0.5' means that the classifier is rather poor at separating the pair of accents. The distribution of coefficients derived from instances of one class overlaps the distribution of coefficients from instances of the other class, so that the classifier performance is not dramatically better than chance. One can interpret '>0' as showing that a classifier can provide some useful separation between the pair of accents: better than chance. This happens if either (a) *K* is near 0.5, (b) one of the two classes is far more frequent than the other so the probability of getting the answer right by chance is large, or (c) that there is not enough data for that pair of accents to draw a stronger conclusion. (Note that $K>0.5$ implies $K>0$, but $K<0.5$ does not.) A blank box means that the classifier performance was sufficiently variable from one training/test split to another than no conclusions can be drawn, presumably due to (b) or (c) above.

Table 10 shows that many of the accents are sufficiently distinct so that a listener might usefully classify them after listening to a single instance. The classifier performance is shown to be significantly better than chance for seven of the 21 pairs. Table 10 also shows that four of the accents that have distinct means are not separable. As with the MANOVA results, the classifiers did not distinguish among several varieties of final rise: the L*H,%, the L*H,H% and the L*,H%. Beyond that, the final boundary tone could not be distinguished between H*L,% and H*L,H% or between L*H,%

Table 10

Summary of *K*-values for each accent pair

	L*,H% (12)	L*H,H% (32)	L*H,% (187)	L*H,L% (9)	H*L,H% (41)	H*L,% (414)	H*,H% (15)
H*,H% (15)				>0	>0	>0	
H*L,% (414)		>0*	**>0.5***	**<0.5***	**<0.5***		
H*L, H% (41)		>0	>0*				
L*H,L% (9)			**<0.5**				
L*H,% (187)	**<0.5***	**<0.5***					
L*H,H% (32)	**<0.5***						
L*,H% (12)							

and L*H,L%. Finally, L*H,L% could not be distinguished from H*L,%. Given that the means are distinct, the low *K*-values for these pairs implies that the acoustic variability of those accents is large enough so that they overlap strongly, at least in terms of F0.