Quantitative Modelling of Intonational Variation

Esther Grabe, Greg Kochanski and John Coleman
Phonetics Laboratory
University of Oxford, United Kingdom

1. Introduction

Intonation is affected by a large number of factors. Among these are dialect and utterance type. In this paper, we investigated the effects of dialect on the intonation of statements and questions. Our research methods were computational–statistical and we concentrated on fundamental frequency, the primary acoustic correlate of intonation. We hypothesised (1) that within dialects, the question/statement distinction would affect the shape of f_0, (2) that dialect would have an effect on the distinction and (3) that we would find evidence of a putative intonational universal: speakers are said to have higher f_0 in questions, compared to statements (Bolinger 1978, Ohala 1983, Gussenhoven 2002, Haan 2002).

Our speech data were taken from the IViE corpus, an existing corpus of recordings from seven urban dialects of English spoken in the British Isles (Grabe, Post and Nolan 2001, Grabe forthcoming). The recordings were made in Belfast, Bradford, Cambridge, Dublin, Leeds, London and Newcastle. The London data were produced by speakers of West Indian descent and the Bradford data by speakers of Punjabi descent. Three male and three female speakers of each dialect read a list of declaratives, wh-questions, yes/no (polar) questions and declarative questions. From these data, f_0 values were extracted and modelled mathematically. In particular, for each f_0 trace in the corpus, we generated orthogonal polynomial models of f_0. In the models, lower coefficients model large scale structures; higher coefficients capture features that change on shorter time scales.

Two findings emerged. Firstly, we found that both dialect and utterance type affected the shape of f_0. We also found that differences in f_0 between questions and statements were made throughout the utterance, in the shape of the contour and in the register. Traditional accounts of English intonation describe questions as having a final rise in f_0 and statements as having a final fall. This account is valid in some dialects, but not in all.

Secondly, we found some common behaviours across dialects: in all dialects, average f_0 was lowest in statements, higher in wh- and yes/no questions and highest in declarative questions. This observation has been made for a number of other languages and it may be evidence of an intonational universal.

We also found that in all dialects, f_0 sloped downwards in declaratives. Declarative questions were modelled as level or overall rising. In wh- and yes/no questions the slope did not contribute to the distinction between questions and statements.

2. Background

Intonation systems in the British Isles vary considerably. Studies of Belfast English, for instance, show that Belfast declaratives are produced with rising intonation. In many other varieties of English and in the so-called standard, declaratives fall (e.g. Jarman and Cruttenden 1976, 1995, Rahilly 1991, Wells and Peppé 1996, Lowry 1997). The intonation of English spoken in Tyneside has been studied by Pellowe and Jones (1978) and by Local, Kelly and Wells (1986). Liverpool English has been investigated by Knowles (1978). Wells (1982), Tench (1990) and Walters (1999) described the intonation of Welsh English. London Jamaican has been investigated by Sebba (1993) and by Sutcliffe and Figueroa (1992). The intonation of Glasgow English has been described by Mayo, Aylett and Ladd (1996) and by Vizcaino-Ortega (2002). The intonation of English spoken in Manchester has been investigated by Cruttenden (2001).

Multi-dialect comparisons, however, are rare; the collection of comparable speech data from a number of dialects requires time and resources. Comparison of the intonation of several German dialects were carried out by and by Auer, Gilles, Peters and Selting (2000) and by Ulbrich (2002). Some comments on intonation systems in dialects of English can be found in Wells (1982). Between 1997 and 2002, a research project on intonational variation in English was conducted at the University of Cambridge. During the IViE project[1], a machine-readable corpus of speech data was collected, designed specifically for the investigation of intonational variation in the British Isles. The design included four factors that affect intonation: dialect, speaking style, speaker and gender. The completed corpus contained 36 hours of directly comparable speech data from seven urban dialects of English spoken in the British Isles. In local schools, twelve sixteen-year-old speakers from each dialect produced data in five speaking styles: read sentences, read text, semi-spontaneous recall of the read text, goal-directed interaction and free conversation. The data were digitised, catalogued and made publicly available[2]. Five hours of recordings were annotated orthographically and prosodically using a two-tone (HL) prosodic labelling system initially based on ToBI (Silverman et al. 1992, Beckman and Ayers 1997) and later adapted for the transcription of intonational variation in the British Isles (Grabe, Nolan and Farrar 1998, Grabe, Post and Nolan 2001, Grabe 2002).

Two findings emerged (Grabe, Post, Nolan and Farrar 2000, Grabe and Post 2002, Grabe forthcoming). Firstly, we found that the intonational differences between some English dialects can be as great as intonational differences between two languages, e.g. between English and German (Grabe 1998). Secondly, we showed that intonational variation is more considerable than textbooks on English intonation suggest (e.g. O'Connor and Arnold 1973). Considerable variation was observed between dialects, within dialects, between speakers and within speakers. On identical texts and in identical contexts, speakers from the same dialect produced a number of different intonation contours. Between-dialect differences involved the usage and frequency of contours, not specific contour shapes and distributions overlapped across dialects and speakers. We

concluded that current models of intonation could not account very well for the variation in our data.

The effect of utterance type on intonation was investigated by Grabe (2002). Grabe examined the recordings of read sentences in the IViE corpus. These contained a variety of syntactic structures, among them declaratives (*You remembered the lilies.*), wh-questions (*Where is the manual*?), yes/no questions (*May I lean on the railings*?) and declarative questions (questions without morphosyntactic question markers, *You remembered the lilies*?). Data from three male and three female speakers per dialect had been intonationally labelled (714 intonation phrases) during the project. The intonation labels were subjected to statistical analysis. The results showed that dialect affected the realisation of the question–statement distinction but they also revealed cross-dialect similarities. In all dialects, a final rising contour was significantly more likely if a question contained fewer syntactic or morphological question cues. Similar effects have been observed in single dialects of other languages (Grabe and Karpinski 2003: Polish; Haan and van Heuven 1999, Haan 2002: Dutch; Brinckmann & Benzmüller 1999: German).

In the present paper, we add an acoustic investigation of the same set of 714 sentences. The study summarised above was restricted to the incidence of final rises. In the present study, we explored the distinctions between the four utterance types throughout the utterance.

3. Method

The list of sentences is given in Appendix A[3]. They were read by three male and three female speakers from each of seven dialects.

Our analysis was conducted on three measures derived from the acoustic data:

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

In the test sentences, we analyzed fundamental frequency, using the loudness and periodicity signals to weight the importance and reliability of different regions. This approach takes into account the observation that not all the f_0 measurements in an utterance are equally important or reliable. We are concerned with the learned, controlled pitch shapes, and are not interested in pitch perturbations near phoneme transitions (except to the extent they may be used with intent to communicate). Secondly, regions of speech where the value of f_0 is not clear are unimportant, because those regions do not help us to separate a good model of the intonation from a bad one. Thirdly, one would expect the importance of an f_0 data point to decrease as the loudness decreases.

---

[3] The test sentences form part of the IViE corpus and they can be downloaded from http://www.phon.ox.ac.uk/~esther/ivyweb/download1.html. Note that the corpus contains data from 12 speakers per dialect; the six speakers chosen for the purposes of the present study are the speakers whose data was prosodically annotated.

Combining these considerations, we assigned a weight W(x) to each data point, as described in Appendix B [B.1]. Use of such a weight focuses our mathematical attention on the most reliable, important parts of the utterance, and will give us an analysis that better represents what would be important to a human listener.
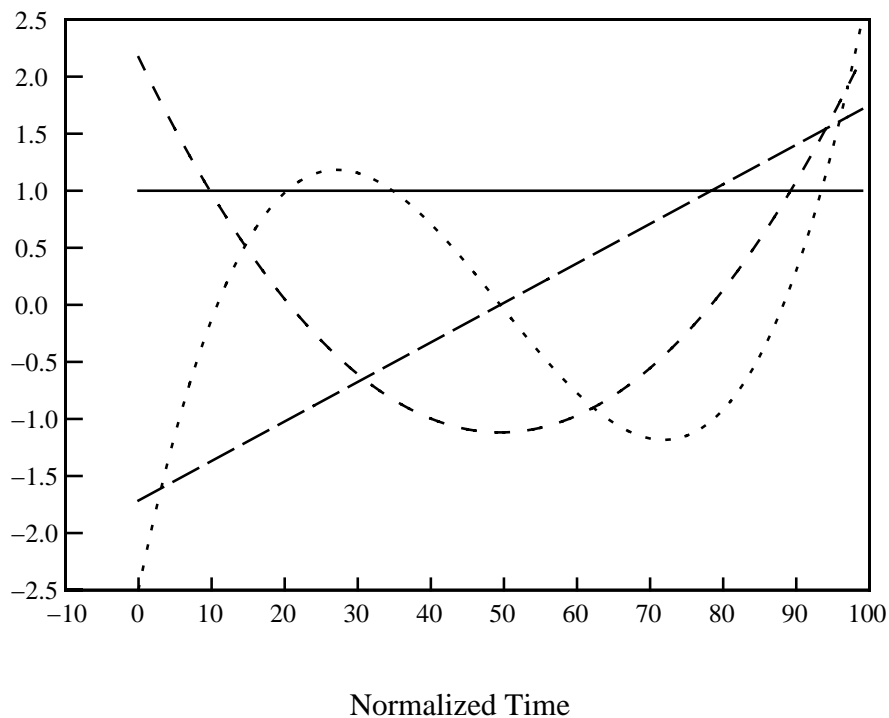

Normalized Time

Figure 1. The figure shows the lowest four Legendre polynomials, the first (solid), then second, third, and fourth with successively shorter dashes.

Before the data were analyzed, it was inspected for gross errors in the f_0 tracks. An automated procedure was run to identify likely problem areas, and then a human labeller (the first author) inspected the area and sometimes marked a change.

We defined the analysis to cover only the voiced region of the intonational phrase. In the section of the IViE corpus investigated here, all utterances were designed to be fully voiced. 12 utterances per dialect, however, ended in a voiced fricative (*They are on the railings*). Here voicing was not as regular as in other sections of the utterances. The decision was also relevant to a small number of other utterances that ended in irregular voicing.

The central step in the data analysis was to represent the data as a best-fit sum of Legendre polynomials with each polynomial normalised to have unit variance (Figure 2). The result of the analysis is a model for the f_0 of each utterance. The model is a somewhat smoothed version of f_0 that bridges over unvoiced regions (see Figure 2).

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

$$M(x) = \sum_i c_i \cdot f_i(x) \ .$$
(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. The Nth Legendre polynomial picks out variations in f_0 which have a scale of *2/N* of an intonational phrase.

The analysis of the data is performed by a weighted linear maximum-likelihood regression, after the pitch was normalized both in time and normalized in frequency. All the pitch curves were scaled to the same length, and all were divided by the speaker's mean pitch. (See Appendix B for details.)
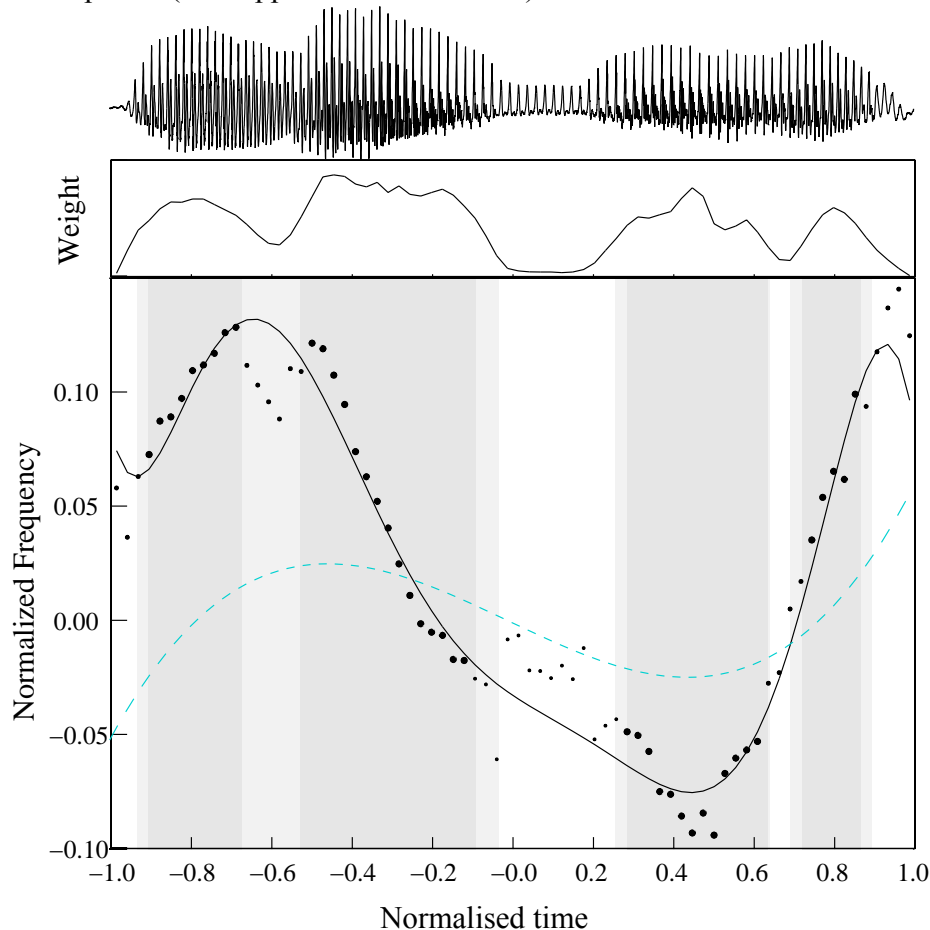

Figure 2. Illustration of the analysis procedure for a typical utterance. Top : acoustic signal; middle: weight for fit to f_0; bottom: f_0 data (dots), fit (solid line), and the contribution of the fourth Legendre polynomial (dashed line).

Figure 2 shows that the acoustic signal (top) is analyzed to yield f_0, loudness and voicing measurements. The loudness and voicing measurements are combined to yield a weight (middle) which will be used to control the regression. The bottom panel shows the f_0 data (dots), and the best fit model that comes from the regression analysis (black line). One of the larger terms in the sum (Eq. 1) is shown in dashed grey. This is the fourth Legendre polynomial times the corresponding coefficient determined by the regression analysis. When eight such terms are added together, the black curve results. The grey areas show

where the weight is large (dark grey: weight>50% of maximum, light grey: weight>25% of maximum). In these regions, the model is forced to match the f_0 data most precisely. If one coefficient from the regression is particularly large, the data and the model will tend to have the shape of the corresponding orthogonal polynomial.

The first few coefficients have straightforward physical interpretations:

1. The first coefficient, c_0, is just the average f_0 of the utterance after normalization by equation B.2 in the Appendix. So, if $c_0 = 0.1$, for example, the utterance's average f_0 is 10% higher than that speaker's average. Note that this coefficient reveals utterance-to-utterance differences. In normal conversation, humans should be able to notice that an utterance is higher pitched than normal for the person who is talking. We note that this is a major difference from many techniques for representing intonation, such as ToBI (Beckman and Ayers Elam 1997), which are normally applied to a single utterance and therefore cannot consider the kind of overall contrast that this coefficient reveals.
2. The second, c_1, is half the best-fitting slope of the utterance, expressed as a fraction of the speaker's average f_0 over the utterance. So, $c_1 = -0.05$ corresponds to a modest (10%) decline in f_0 over the utterance. If $c_1 = 0$, there is no declining trend to an utterance. (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 centre of the utterance. An utterance with no overall curvature would have $c_2 = 0$.
4. Succeeding terms correspond to features of successively shorter duration. For these sentences, which average 4.6 ± 1.3 words long, with 1.3 ± 0.5 syllables per word, coefficients in the region c_4 to c_7 would respond to f_0 bumps on the scale of a word or accent, c_5 to c_9 would be most sensitive to bumps on the scale of a single syllable, and even higher coefficients would correspond to changes in f_0 shorter than a syllable.

4. Results

The results section contains the following selection of findings. Firstly, we show reconstructions of typical f_0 traces for each utterance type in each dialect, using c_0–c_7.
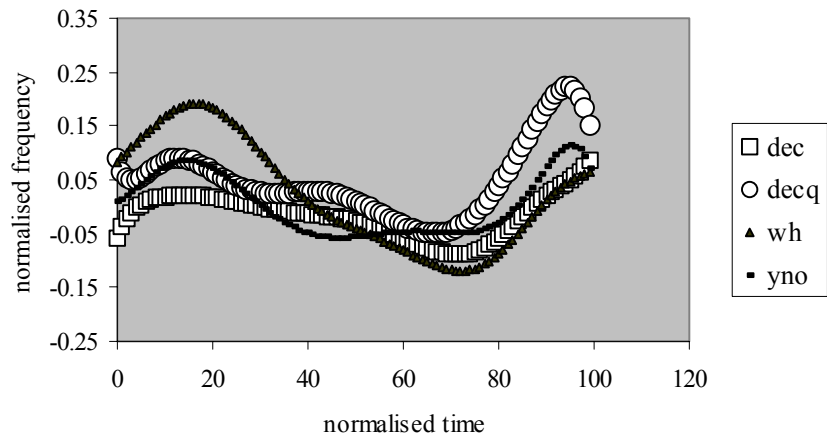
Secondly, we show that the first two coefficients (average and slope) contribute strongly to the distinction between the four utterance types. The contribution of higher coefficients is marginal.

Thirdly, we illustrate the distribution of average and slope values for the four utterance types in the seven dialects investigated. In all dialects, average and slope distinguish declaratives from declarative questions (but note that the f_0 patterns produced in the utterance types differed across dialects, cf. Figures 3–9 below). Wh- and yes/no questions were intermediate.
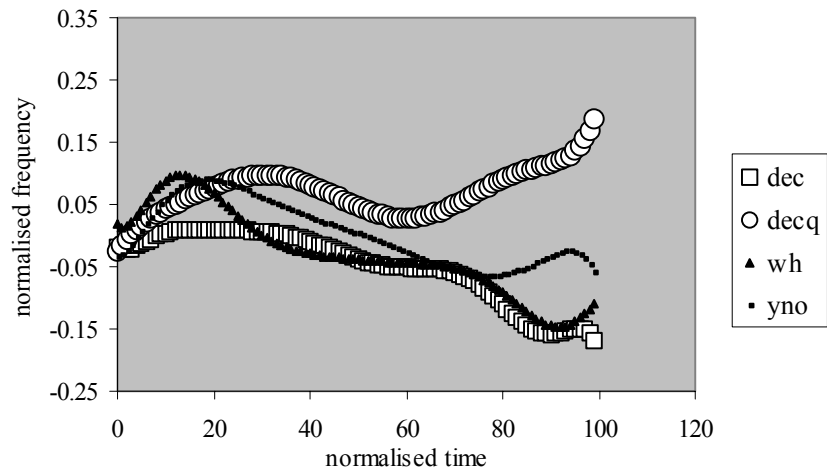
Figure 3–9 show typical f_0 contours for each utterance type in each dialect. The contours were constructed by taking the median value for each coefficient across all

utterances of a given type in a given dialect, then computing f_0 from these median coefficients via Equation 1. (Note that our speakers also produced a range of other contours, particularly in nuclear position. A linguistic analysis of variation in nuclear position is given in Grabe 2002.) In the graphs, normalised time is shown on the *x*-axis. Normalised frequency is shown on the *y*-axis. Unfilled circles and squares represent statements and declarative questions. As one would expect from earlier work on dialect intonation in English, the graphs show that typical f_0 contours differ between dialects.
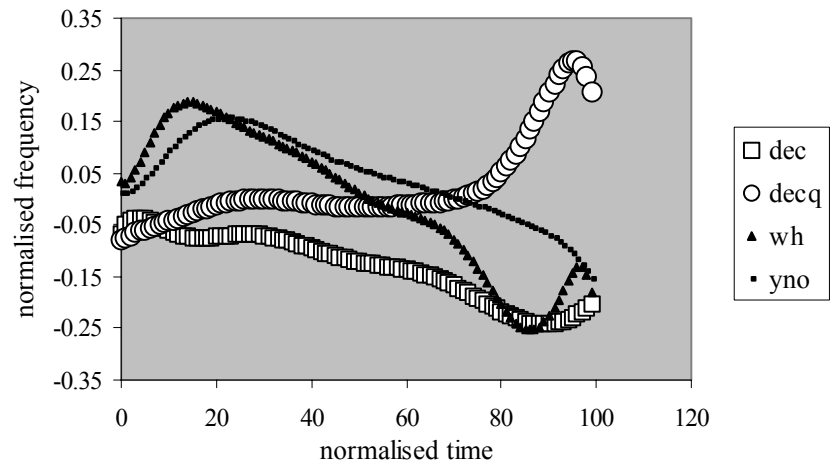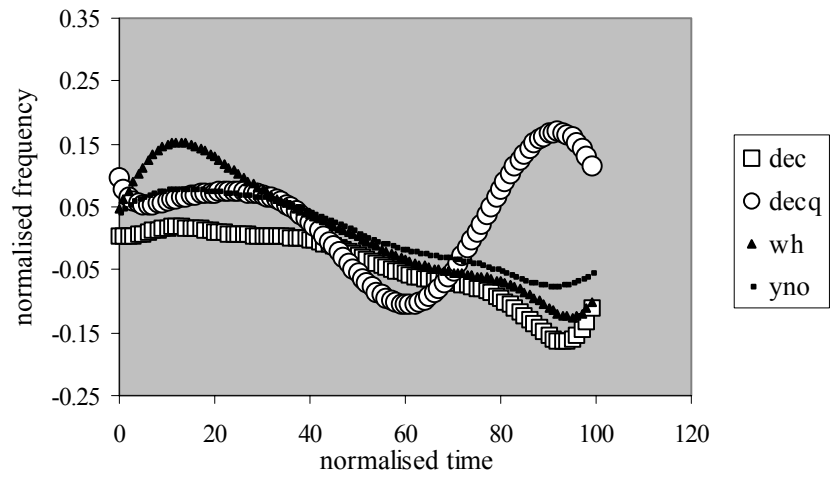
Figure 3: Belfast
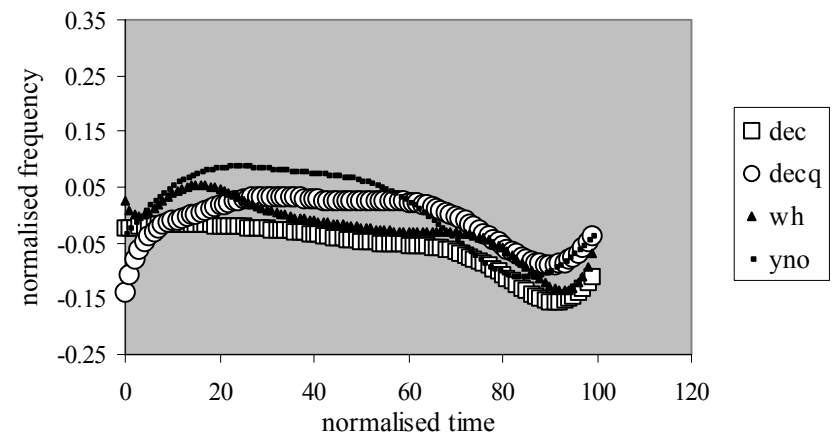

Figure 4: Bradford

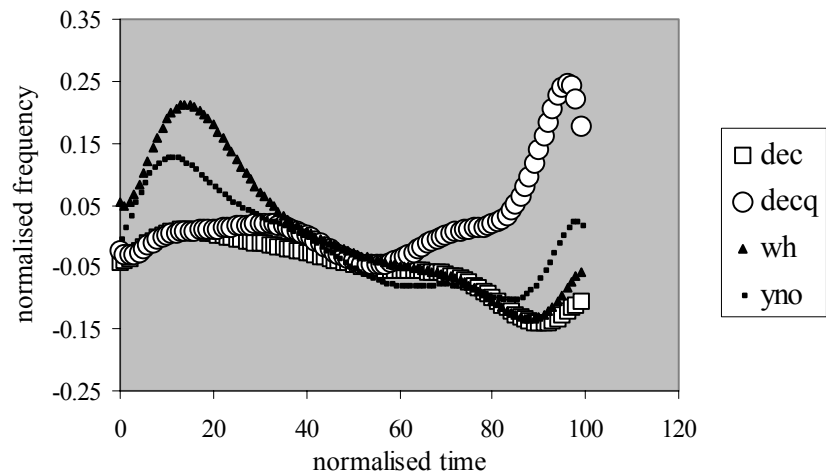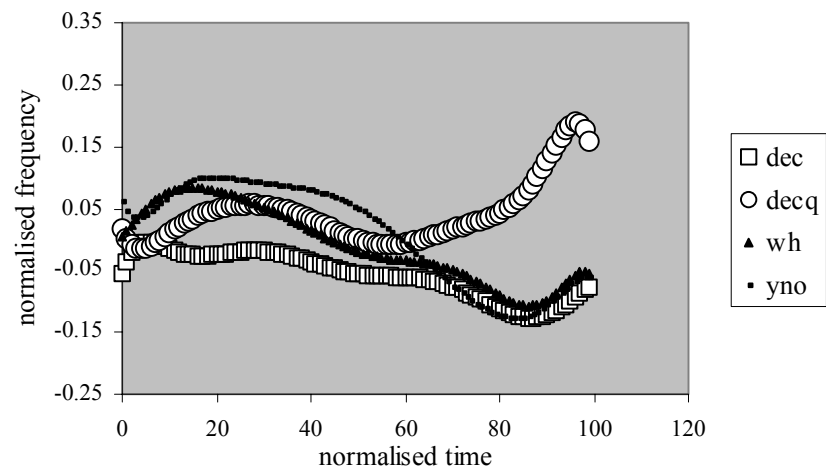**Figure 5: Cambridge**

Figure 6: Dublin

Figure 7: Leeds

Figure 8: London

Figure 9: Newcastle

Figures 3–9. Median f_0 traces for seven urban dialects of English reconstructed from orthogonal polynomials.

Figures 3–9 shows that in the four utterance types, typical contours differed in Cambridge, London and Dublin. In Belfast and Leeds, the contours were very similar. In Belfast, all utterance types were produced with final rises in f_0[4]. In Leeds, the majority of contours was produced with overall falling f_0 patterns. In Newcastle and in Bradford,

---

[4] Some of the figures show a small downturn in f0 after a high final rise (e.g. Belfast, London, Dublin, Cambridge, declarative questions). Invariably, these were produced with low amplitude and may be perceptually unimportant. Physiologically, fast vocal fold vibration may be harder to switch off than slow vocal fold vibration.

finally, declarative questions were likely to rise. In the other utterance types, f_0 sloped downwards.

Then we examined the contribution of individual coefficients to the distinction between utterance types. Figure 4 shows mean values for the first eight coefficients.

Figure 4. Size of the first eight coefficients of orthogonal polynomial models generated for each utterance type. The data are summed over dialects.

The figure shows that the contribution of the first and the second coefficients (average and slope) is the largest. The higher coefficients contribute successively less to the shape of the utterance and the difference between sentence types. In Figure 5, we show how the first two coefficients differentiate the four utterance types.

Average f0 and f0 slope in seven dialects

Figure 5. Average f_0 (*x*-axis) plotted against the global slope of f_0 (*y*-axis). The figure shows data for four utterance types and seven dialects.

For average f_0, positive values show that average in an utterance type exceeded the mean for the speaker group from the dialect and vice versa. Positive slope values indicate a globally rising slope.

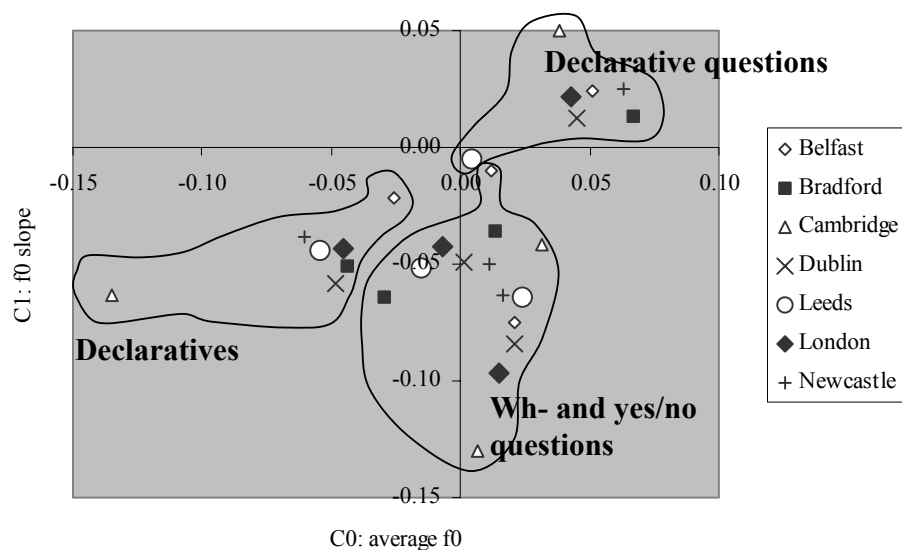The figure shows evidence of substantial cross-dialect differences but also some broad similarities. Typical utterance of each type fall into three distinct groups: declaratives, declarative questions and other questions. (Note that the plotted points are medians and the individual utterances are somewhat more scattered). Within a dialect the points from different groups are well separated even if some of them may be close to points from other dialects. The four utterance types appear to be distinct within every dialect in addition to the cross-dialect patterns marked in the figure.

6. Discussion and conclusion

Our data provide quantitative evidence of dialect differences in English intonation. In some dialects, speakers produced similar-shaped contours in the four utterance types investigated (e.g. in Belfast or in Leeds), in others, the contours were clearly different (e.g. in Cambridge or in London). This suggests that contour type may well support the question/statement distinction, but the presence of such information in the signal is dialect-specific.

We found a consistent correlation between average f_0 and the question/statement distinction. Average f_0 in questions was higher than in statements and the height of the average was affected by the lexical and/or syntactic characteristics of the question. In all dialects, average f_0 was highest in declarative questions, lower in other questions and lowest in declaratives. Comparable observations have been made in many other languages. Herman (1942) showed that pitch was rising or higher in questions and falling or lower in declaratives in a sample of 175 languages. Ultan (1969) and Bolinger (1978) found similar results in samples of 53 and 41 languages, respectively (cf. also Ohala 1983, Haan 2002 and Gussenhoven 2002).

We also observed an effect of utterance type on the slope of f_0, but this effect was not directly comparable to the effect on average f_0. Thorsen (1978) found that in Danish, declaratives were most steeply sloping, declarative questions were horizontal and other questions and non-final clauses filled the intermediate space. In Danish, however, accent pattern is not a linguistic choice; only one pattern is possible. In English, several patterns are possible and in our data, Thorsen's observation was not replicated for any of our dialects. Wh-questions decline more steeply than declaratives and the slopes of yes/no questions do not differ substantially from those of declaratives.

A comparable observation was made by Haan (2002) in her investigation of Dutch, another language in which accent pattern is a linguistic choice. Haan investigated the same utterance types as the ones discussed here and concluded that the slope of f_0 reflected accentuation rather than a linguistic choice of slope on the part of the speaker. English and Dutch wh-questions are characterised by early and prominent accents on the wh-word followed by weaker and frequently downstepped accents on the subject. These accent patterns, Haan concluded, were responsible for the steeply declining slopes in wh-questions.

A similar explanation may account for the declining declaratives and inclining declarative questions in our data. Grabe (2002) showed that in our data, in all dialects, late rises in f_0 ('nuclear rises') were produced most frequently in declarative questions and least frequently in declaratives. Consequently, the slope of a declarative questions is more likely to rise than the slope of a declarative. An additional effect of utterance type on the slope cannot be ruled out, but we cannot disentangle the effects, at least not in our data.

Our findings raise a question. They show that differences in average f_0 could be meaningful. Can we therefore conclude that average f_0 and the accompanying register differences are linguistic? Many of the current models of intonational phonology do not consider that register differences can be linguistic. Since register differences are gradient, they are therefore frequently classed as paralinguistic. In addition, some authors argue that high f_0 in questions is universal and biological in origin (Ohala 1983, Gussenhoven 2002). Some people would therefore assume that the effect is not part of linguistic structure. Our data broadly supports the hypothesis that high f_0 in questions is universal but we see no reason consider that because a behaviour is biologically defined it cannot be co-opted by the language faculty for linguistic purposes.

Ladd (1996: 272–277), for instance, argues that register differences in downstep can be phonological. Ladd's proposal is based on the notion of relative strength in metrical phonology. He argues that downstep can be modelled as a syntagmatic relation of pitch level between two accents or other prosodic constituents such as intonation phrases[5]. A raised register in questions can be modelled similarly, as part of the phonology of intonation. However, raised registers may also be meaningful in isolated utterances (imagine someone shouting *watch out*! at the top of her voice)[6]. Consequently, a syntagmatic model of register differences is necessarily partial.

In conclusion, our data show that in seven dialect of English dialects, a raised average f_0 accompanies questions. In addition, we found evidence of a trade-off between non-prosodic characteristics of the utterance and f_0. Our data do not provide support for the hypothesis that the global slope of f_0 is correlated with a trade-off between lexical and/or syntactic question cues. The slope is, however, correlated with the distinction between declaratives and declarative questions. Finally, the data show that in some dialects (e.g. Cambridge or London), the distinction between our four utterance types involves different accent patterns. In other dialects (e.g. Belfast or Leeds), we did not find evidence of localised differences in f_0.

---

[5] Ladd 1988 and van den Berg, Gussenhoven and Rietveld 1992 showed that downstep can apply across intonation phrases.

[6] Ladd 1996, 252-257 reviews normalising ('syntagmatic') and initialising ('paradigmatic') accounts of the perception of pitch range.

9. References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. New York: Dover Publications, ninth (1970) printing, 773–802.

Auer, P., Gilles, P., Peters, J., Selting, M. (2000). Intonation regionaler Varietäten des Deutschen. Vorstellung eines Forschungsprojekts. In Stellmacher, Dieter (ed.), *Dialektologie zwischen Tradition und Neuansätzen*. Beiträge der Internationalen Dialektologentagung, Göttingen, 19.–21. Oktober 1998, Stuttgart: Steiner, 222–239.

Beckman, M. E. and Ayers Elam, G. (1997). Guidelines for ToBI Labelling, version 3. The Ohio State University Research Foundation, Ohio State University.

Bel, B. and Marlin, I. (eds.), *Proceedings of the Speech Prosody 2002 Conference*, 11–13 April 2002, Aix-en-Provence: Laboratoire Parole et Langage,

Bolinger, D. (1978). Intonation across languages. In Greenberg, J. (ed.), *Universals of Human Language. Vol. 2: Phonology,* Stanford University Press, 471–524

Brinckmann, C. and Benzmüller, R. (1999). The relationship between utterance type and F_0 contour in German. In *Proceedings of Eurospeech 1999*, Vol. 1, 21–24.

Grabe, E. (2002). The IViE Labelling guide. http://www.phon.ox.ac.uk/~esther/ivyweb/guide.html.

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

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

Gelman, A., Carlin, J. B., Stern, H. S. and D. B. Rubin (2000). *Bayesian Data Analysis*. London: Chapman and Hall, 7–21 and 78–87.

Grabe, Esther (1998). Comparative Intonational Phonology: English and German. Grabe, E. 1998a. *Comparative intonational phonology: English and German*. Max Planck Institute for Psycholinguistics Series 7, Wageningen, Ponsen et Looien

Grabe, Esther (2002). Variation adds to prosodic typology. In B. Bel and I. Marlin (eds), 127–132.

Grabe, E. (forthcoming) Intonational variation in English. In Gilles, P. and Peters, J. (eds). *Regional Variation in Intonation.* Linguistische Arbeiten, Niemeyer.

Grabe, E. and Karpinski, M. (2003). Universal and language-specific aspects of intonation in English and Polish. In *Proceedings of the 15th International Congress of Phonetic Sciences*, 3–9 August, Barcelona, Vol. 1, 1061–1064.

Grabe, E. and Post, B. (2002). Intonational Variation in English. In B. Bel and I. Marlin (eds), 343–346.

Grabe, E., Post, B. and Nolan, F. (2001). Modelling intonational variation in English: the IViE system. *Proceedings of Prosody 2000*, Adam Mickiewicz University, Poznan, Poland, 51–58.

Grabe, E., Brechtje P., Nolan, F. and Farrar, K. (2000). Pitch accent realisation in four varieties of British English. *Journal of Phonetics* 28, 161–185.

Grabe, E., Nolan, F., and Farrar, K. (1998). IViE — A Comparative Transcription system for Intonational Variation in English. *Proceedings of the 5th Conference on Spoken Language Processing (ICSLP)* 1998.

Grabe, E., Post, B. and Nolan, F. (2001). *The IViE Corpus*. Electronic resource available from http://www.phon.ox.ac.uk/~esther/ivyweb.

Gussenhoven, C. (2002). Intonation and interpretation: Phonetics and Phonology. In B. Bel and I. Marlin (eds), 47–57.

Haan, J. (2002). *Speaking of questions*. Utrecht: LOT.

Haan, J. and van Heuven, V. (1999). Male vs. Female pitch range in Dutch questions. In *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, Vol. 2, 1581–1584.

Herman, E. (1942). *Probleme der Frage*. Nachrichten von der Akademie der Wissenschaften in Göttingen, Philologisch-Historische Klasse, Nr. 3, 4.

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

Knowles, G. O. (1978). The nature of phonological variables in Scouse. In P. Trudgill (ed.). *Sociolinguistic patterns in British English*. London: Edward Arnold.

Kochanski, K. and Shih, C. (2003) Prosody Modeling with Soft Templates. *Speech Communication* 39, 311–352.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge, CUP.

Ladd, D. R (1988). Declination 'reset' and the hierarchical organization of utterances. *Journal of the Acoustical Society of America* 84, 530–44.

Local, J., Kelly, J. and Wells, W. (1986). Towards a phonology of conversation: turn-taking in urban Tyneside speech. *Journal of Linguistics* 22: 2. 411–437.

Lowry, O. (1997). *Intonation patterns in Northern Irish English*. Unpublished MPhil Thesis: University of Cambridge.

Mayo, C., Aylett, M. and 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*, 231–234.

O'Connor, J. D. and Arnold, G.F. (1973). *Intonation of colloquial English.* London: Longman.

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

Ohala, J. and Ladefoged, P. (1970). Further Investigation of Pitch Regulation in Speech. *UCLA Working Papers on Phonetics,* 14, 12–24.

Pellowe, J. and Jones, V. (1978). On intonational variability in Tyneside speech. In P. Trudgill (ed.) *Sociolinguistic patterns in British English*. London: Arnold.

Press, W. H. Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge: Cambridge University Press, 671–680.

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

Sebba, M. (1993). *London Jamaican: language systems in interaction.* London: Longman.

Silverman, K., Beckman, M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J. and Hirschberg J. (1992). ToBI: a standard for labeling English prosody. In *Proceedings of the 1992 International Congress of Spoken Language Processing*, Banff, Canada, Vol. 2, 867–70.

Stevens, S. S. (1971). Perceived Level of Noise by Mark VII and Decibels (E). *Journal or the Acoustical Society of America*, Vol. 51 (2 part 2), 575–602.

Sutcliffe, D. with J. Figueroa (1992). *System in Black Language*. Clevedon, Avon: Multilingual Matters.

Tench, P. (1990). The pronunciation of English in Abercrave. In N. Coupland (ed) *English in Wales*. Clevedon, Avon: Multilingual Matters.

Thorson, N. (1978) An acoustic analysis of Danish intonation. *Journal of Phonetics* 6, 151–175.

Ulbrich, C. (2002). A comparative study of intonation in three standard varieties of German. In B. Bel and I. Marlin (eds), 671–674.

Ultan, R. (1978). Some general characteristics of interrogative systems. In J. Greenberg (ed), *Universals of human language*, *Vol. 4: Syntax.* Stanford University Press. 211–248.

van den Berg, R., C. Gussenhoven and T. Rietveld (1992). Downstep in Dutch: Implications for a model. In G. J. Docherty and D. R. Ladd (eds.) *Papers in Laboratory Phonology II. Gesture, Segment, Prosody*, Cambridge: Cambridge University Press, 335–58.

Vizcaino-Ortega, F. (2002). A Preliminary Analysis of Yes/No Questions in Glasgow English. In B. Bel and I. Marlin (eds.), 683–6.

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*, second edition, John Wiley and Sons, New York, 80–105.

Wells, B. and Peppe, S. (1996). Ending up in Ulster: prosody and turn-taking in English dialects. In E. Couper-Kuhlen, M. Selting, *Prosody in Conversation: Interactional studies.* Cambridge: Cambridge University Press

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

Xu, Y. (2000) How fast can we really change pitch? Maximum speed of pitch change revisited. In *Proceedings of the Sixth International Conference on Spoken Language Processing,* Beijing, 16–20 October 2000.

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 f_0,
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 f_0 signal. We will first describe how these measures were derived.

We computed a binary voicing estimate and an estimate of the fundamental frequency, f_0, using the *get_f0* program from the ESPS package (Entropic Corp.). As described below, f_0 was checked, perhaps adjusted, then normalized.

We used an approximation to the steady-state perceptual loudness. We used Stevens' Mark VII computation (Stevens 1971), modified slightly to use 0.7 octave frequency bins rather than the full octave bands or one-third octave bands for which it is originally defined. It operates on the spectral power density derived from a 50 ms wide $1 + \cos(\pi \cdot \dfrac{t}{25\,\mathrm{ms}})$ window. Stevens' method is an improved version of the ISO-R532 Method A standard noise measurement. Before use, we subtracted the background noise from the loudness estimate.

The periodicity measure ranges from 0 to 1, with zero assigned to regions where the signal cannot be predicted, and one to regions where the signal is locally an exact copy of the past. The audio was pre-processed, with a 50 Hz fourth-order time-symmetric Butterworth high-pass filter to eliminate DC offsets and low-frequency noise. Then, the periodicity measure was derived by taking a 20 ms section of the filtered signal and comparing it to 20 ms sections that ended 2 to 20 milliseconds in the past. If the acoustic signal were exactly periodic with f_0 between 50 and 500 Hz, then one of the past windows would exactly match the signal, and the difference would be zero. The periodicity measure is then one minus the minimum variance of the difference (the minimum is taken over past windows) divided by the local variance of the signal. The result is similar to taking the maximum cross-correlation between windows.

Weighting of the Data

Not all the f_0 measurements in an utterance are equally important or reliable. For instance, the first cycle or two of the glottal oscillation are often quite different from the remainder of the voiced region. This is best explained as intrinsic behaviour of the glottal oscillator, rather than anything intentional. To reflect this, we automatically reduced the weight of the first and last 10 ms of data by a factor of 4 in every voiced region ('edge reduction').

A second case where the weight should be low is in places where the speech signal is not very periodic. Even assuming that there is a single true f_0 for such signals,

get_f0 is less likely to find it, given the weak acoustic evidence. Thus, the resulting f_0 estimate will be unreliable.

Further, *get_f0*, like many f_0 tracking algorithms, internally assumes a Markov Chain model for f_0, and it will interpolate across places where the acoustic signal is not very periodic. While this is fine for many purposes, it means that f_0 in weakly periodic regions is not an independent measurement: it can be thought of as a copy of the nearest region where f_0 is well determined. Including the periodicity estimate in the data's weight function handles all this cleanly. All other things being equal, the periodic regions at the centre of vowels will count more in further analysis, and the less-periodic regions at phone transitions or in irregular voicing will be properly de-emphasized.

Third, low weights are appropriate in quiet regions. One expects that the importance of an f_0 datum will increase as the loudness increases. Certainly, under normally noisy conditions, the parts of an utterance that are more than 10 or 20 dB quieter than the peaks (loudness levels 0.4 to 0.2 times the peak loudness) will often be covered up by ambient noise and therefore have no perceptual importance. Even neglecting noise, the f_0 of quiet regions of speech seems less important to us: for instance, in the tail of *Ealing* (/ˈiːlɪŋ/, an area of London) at the end of an utterance. Such utterances often have a rise or fall well after the centre of the last syllable, once the acoustical power is 20 or more decibels down. The presence or absence of this quiet excursion can be entirely unsuspected if one only listens to the speech. Thus, it seems clear that the weight of a datum should be proportional to some function that gives zero weight to silence, and increases as the loudness increases.

Before the data was analyzed, it was inspected for gross errors in the f_0 tracks. We also identified regions of irregular phonation that did not have a clear pitch, and reduced their weight in Equation B.1 by setting $H(x) = 0.25$ (*vs.* 1 normally). An automated procedure was run to identify likely problem areas, and then a human labeller inspected the area and sometimes marked a change. In the section of the corpus used here, 254 utterances were marked, with a total of 498 regions marked, of which 75 regions were octave shifts of f_0, and the vast majority of the remainder were either marked as having irregular phonation or no phonation. The median length of the marked regions was 0.056 seconds long.

Combining all these considerations, we assign a weight of

$$W(x) = E(x) \cdot P^2(x) \cdot L^2(x) \cdot H(x) \qquad \text{[B.1]}$$

to each datum, where *E(x)* is the edge reduction, *P(x)* is the periodicity measure, *L(x)* is the loudness, and *H(x)* is the result of hand marking of edges and irregular phonation. This product form for the weight is somewhat arbitrary, though it has the advantage that it assigns a large weight only when the acoustic signal is loud and has a well-defined pitch.

Scope of the Analysis

We defined the analysis to cover only the voiced region of the intonational phrase. This decision was primarily relevant to twelve utterances per dialect which end in a fricative (*They are on the railings*), but also had some effect on a few other utterances that just

trailed off or developed irregular voicing at the end. This choice is equivalent to saying that, if one were to convert the tail of an utterance from voiced to unvoiced, the intonational contour would compress into the remaining voiced region. The alternative choice, fitting the orthogonal polynomials to the entire utterance, is equivalent to the hypothesis that de-voicing the tail would truncate the intonation contour.

To test this fricative compression hypothesis, we looked at twelve of the *railings* sentences over three dialects and compared them to fully-voiced utterances of the same type. We saw final rises and falls that showed evidence of truncation in only one of the twelve cases. The other eleven pairs showed matching intonation patterns and supported the hypothesis that the intonation pattern would be compressed as the voiced region shrinks.

The ends were trimmed off an utterance to the extent that the weight function [Eq. B.1] was less than 1% of the utterance's average weight function. This resulted in the removal of 24±24 ms from the beginning of a typical utterance and 55±72 ms from the end. For the six "railings" sentences per dialect, 31±21 ms were removed from the beginning and 160±60 ms from the end.

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:

1. They can be used, reversibly, to analyze a curve and to reconstruct it.
2. 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.
3. They are orthogonal, which means that, in a specific sense, they do not overlap. 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 polynomials 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.

Mathematically, a family of functions is orthogonal if and only if $\sum_x f_i(x) \cdot f_j(x) \cdot w(x) = 0$ for any two different functions, i and j, in the family (*i.e.* $i \neq j$). The sum is taken over the normalized time values [Eq. B.3] over the scope of the analysis. Strictly speaking, one says that the family is orthogonal over the set of x under weighting function $w(x)$. (Note that $w(x)$ and $W(x)$ aren't the same: $w(x)$ is the weight function that defines the family of orthogonal functions, $W(x)$ is the weight function used to analyze a particular utterance.) Since we had f_0 data at a 10 ms frame rate, and a typical utterance was about 3 seconds long, typical x-values might be -1, -0.997, -0.994,

…, 0.991, 0.994, 0.997, 1.000.   We normalized all the orthogonal functions used in our analysis to have unit variance, so that the coefficients that would later result from the analysis could be directly comparable. This adds the additional constraint $\sum_x f_i(x) \cdot f_i(x) \cdot w(x) = 1$.

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 believe that the IViE labels capture much of the information that is both localized and expressible as a binary high/low contrast. Consequently, we selected a global description for this work, to see what the IViE labels might have missed.

We chose a smooth and continuous family of orthogonal polynomials, as we expect that the intentionally controlled aspects of intonation should be, by and large, smooth and continuous (Kochanski and Shih 2003, see especially section 1.2) because f_0 is controlled by muscle tensions which are smooth functions of time. Additionally, we chose a family of polynomials that have a uniform weighting function across the whole utterance: $w(x) = 1$. The weighting function of a family specifies what parts of the utterance the analysis is most sensitive to, and we did not wish to bias the analysis towards any region of the utterance.

As a result, we chose the family known as Legendre polynomials (Abramowitz and 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 nth 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). We used a Bayesian prior that tended to minimize the sum of the squares of all the coefficients. Its strength was such as to reduce well-determined coefficients by 1%, and it (intentionally) would keep small the values of coefficients that were poorly determined by the data of a given utterance. Descriptions of the method can be found in Weisberg (1985) and in Press, Teukolsky, Vetterling, and Flannery (1992). For a description of the implementation of MAP regression (also known as 'Linear Regularization') in terms of a standard linear regression, see (Press et al 1992, 808-813). For a discussion of MAP regression, see Gelman, Carlin, Stern and Rubin (2000). The regularization/MAP analysis was introduced early in the work, before we assumed that the 'railings' utterances would compress. At that point, it was rather important, as some of those utterances are missing f_0 data for a long stretch at the tail and a conventional maximum likelihood approach gave large and correlated uncertainties for the coefficients derived from some utterances. When we introduced the compression hypothesis, the long stretches without f_0 data are removed, and the difference between MAP and maximum likelihood analyses became fairly small.

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 Nth Legendre polynomial picks out variations in f_0 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 f_0 curve, $f(t)$, in terms of that family by standard techniques of multiple linear regression. First, $f(t)$ is normalized to compensate for inter-speaker differences:

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

where $\bar{f}$ is the speaker's f_0, averaged over all the single-speaker utterances in the IViE database. A normalized f_0 of 0.1 corresponds to an f_0 that is 10% above the speaker's average.

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

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

where t_c is the center of the intonational phrase and L is the phrase's length.

Third, a set of equations is determined, one for each f_0 measurement. These equations are a model for the f_0, written as a sum of the orthogonal functions, each multiplied by a constant[7]:

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

where c_i is the (as yet unknown) coefficient that shows how much the ith function contributes to the shape of the f_0 curve. The sum is taken over the first *N* functions in the family. 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 computes the total error for each model and chooses the model that minimizes the error. The error is

$$\chi^2 = \sum_{X} \left(M(x) - F(x) \right)^2 \cdot W(x) + \lambda \cdot \sum_{i} c_i^2, \qquad\qquad\qquad [B.5]$$

where *X* is the set of places where we have f_0 measurements, and *W(x)* controls how much weight we give to errors in different places. The right-hand sum is the Bayesian prior that states that we don't expect the coefficients to be excessively large, and $\lambda = 0.01^2 \cdot \sum_{X} W(x)$.

This total error can tell you which model is the best representation for the observed f_0. Bearing in mind that each combination of coefficients gives you a different

---

[7] Note that this is the same as Equation 1.

model, what we are doing is computing chi-squared for each possible model, and 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.

Determining N (in e.g. Equation B.4) is not always a simple task: if N is too large, the function is "over-fit" and the resulting coefficients can become large and very sensitive to small changes in $F(x)$ or the choice of the Bayesian prior. We set $N = 1 + L/100\,\text{ms}$, allowing enough orthogonal polynomials to put in one complete oscillation every 200 ms which is the maximum rate at which humans can cycle their f_0 up and down (Xu 2000).

The result of the analysis is a set of coefficients, c_0, c_1, c_2, … . The coefficients allow you to reconstruct the data, 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.